FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Predicting Adverse Effects of Drugs

Sofia Alexandra Machado Novais

Mestrado em Engenharia Biomédica

Supervisor: Rui Camacho

July 7, 2017

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MEB - MESTRADO EM ENGENHARIA BIOMÉDICA **2016/2017**

A Dissertação intitulada

"Predicting Adverse Effects of Drugs"

foi aprovada em provas realizadas em 07-07-2017

o júri

Presidente Prof. Doutor Jorge Alves da Silva
Professor Auxiliar do Departamento de Engenharia Informática da FEUP - U.Porto

Prof. Doutor Carlos Manuel Abreu Gomes Ferreira
Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Prof. Doutor Rui Carlos Camacho de Sousa Ferreira da Silva
Professor Associado do Departamento de Engenharia Informática da FEUP - U.Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Autor - Sofia Alexandra Machado Novais

Faculdade de Engenharia da Universidade do Porto

Predicting Adverse Effects of Drugs

Sofia Alexandra Machado Novais

Mestrado em Engenharia Biomédica

Faculdade de Engenharia da Universidade do Porto

July 7, 2017

Resumo

A quantidade de medicamentos que se encontra disponível actualmente, a nível comercial, é elevada. A importância terapêutica e o proveito destes são indiscutíveis. Contudo, os efeitos desconhecidos de medicamentos individuais e/ou a interacção de efeitos entre medicamentos poderá acarretar consequências graves para a saúde da população. O seu desenvolvimento e produção são realizados sob critérios de protecção e segurança, sendo que qualquer medicamento apenas é aceite para venda, após ter sido sujeito a vários testes pré-clínicos exigentes que confirmem o seu benefício e traduzam eficácia no tratamento para o qual serão prescritos futuramente. Apesar de todos os ensaios necessários realizados que antecedem a comercialização de um medicamento, verifica-se que a sua utilização poderá afectar o corpo do paciente de maneira inesperada por diversos motivos. Deste modo, é essencial manter uma vigilância e monitorização dos medicamentos numa fase pós-comercialização.

A utilização de alguns medicamentos pode revelar-se insegura e representativa de riscos, uma vez que a resposta e interacção da população à tomada dos mesmos difere substancialmente. Verifica-se que, na prática, todas as medidas tomadas na fase pré-comercialização poderão ser insuficientes, uma vez que factores como a idade do paciente, historial clínico, interacção com outros medicamentos poderão estar no ápice destes efeitos indesejados.

A Organização Mundial de Saúde (OMS) define Reacção Adversa a Medicamentos (RAM) como "qualquer efeito prejudicial ou indesejado que se manifeste após a administração do medicamento, em doses normalmente utilizadas no homem para profilaxia, diagnóstico ou tratamento de uma enfermidade". Apesar deste problema advir de tempos passados, considera-se que a RAM constitui, actualmente, uma enorme preocupação na saúde pública. Assim, é imprescindível a partilha de informações, experiências e conhecimento, de modo a que seja possível continuar a realizar estudos intensivos que visam manter, não só a confiança da população na toma dos medicamentos, como também a segurança da saúde e bem-estar da mesma.

Os modelos preditivos de Data Mining podem ser uma valiosa ajuda para mitigar o impacto de RAM desconhecida. A prevenção de erros é uma das principais razões para o uso da tecnologia na prática farmacêutica. Actualmente, conhecem-se vários métodos informáticos, cujo objectivo principal é a previsão de efeitos adversos de medicamentos. Assim, os profissionais de saúde podem ser auxiliados pelo desenvolvimento de procedimentos preditivos porque, com essa previsão, muitos efeitos indesejáveis e desconhecidos sérios, causados pela toma de medicamentos, podem ser evitados antes destes serem comercializados no mercado.

Este relatório faz um apanhado de estudos existentes de detecção e previsão de efeitos adversos de medicamentos e do contributo possível da informática para a mitigação deste problema. Em particular, foi feito um estudo teórico e, posteriormente, foram aplicadas as técnicas de Data Mining para prevenir os efeitos adversos de medicamentos.

Abstract

The number of drugs currently available at the commercial level is quite large. The therapeutic importance and the benefit of these are indisputable. However, unknown effects of individual drugs and/or the interaction of effects between drugs may have serious consequences for the health of the population. Although they are developed and produced under safety and security criteria, and any drug is only accepted for sale after having undergone several demanding pre-clinical tests that confirms its benefit and translate efficacy into the treatment for which it will be prescribed in the future. Despite all the necessary tests performed prior to the commercialization of a drug, it is found that its use can still affect the patient's body in an unexpected for a number of reasons. It is therefore essential to maintain the surveillance and monitoring of drugs at a post-marketing stage.

The use of some drugs may prove to be unsafe and risky, since the response and interaction of the population to their use differ substantially. It is known that in practice and despite all measures taken during the premarketing phase they may be insufficient, since factors such as the patient's age, clinical history and interaction with other medicinal products may be at the apex of these undesirable effects.

The World Health Organization (WHO) defines Adverse Drug Reaction (ADR) as "any harmful or undesirable effect that manifests itself after administration of the drug at doses normally used in man for prophylaxis, diagnosis or treatment of a disease." Although this problem has arisen in the past, it is considered that ADRs are currently a major public health concern. It is therefore essential to share information, experiences and knowledge so that it is possible to continue to carry out intensive studies aimed at maintaining not only the population's confidence in the taking of drugs, but also the health and well-being safety of the same.

A valuable help to mitigate the impact of unknown ADRs can be provided by predictive Data Mining models. The prevention of errors is one of the main reasons for the use of technology in pharmaceutical practice. At present, several computer methods are known whose main objective is the prediction of adverse drug effects. Thus, health professionals can be assisted by the development of predictive procedures because, with this prediction, many unwanted and unknown serious effects caused by drug taking can be avoided before they are marketed.

This report gives an overview of existing studies of detection and prediction of adverse drug effects and the possible contribution of informatics to the mitigation of this problem. In particular, a theoretical study was made and, posteriorly, Data Mining techniques were applied to prevent the adverse effects of drugs.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Rui Camacho. I would like to dedicate it a very special acknowledgement for all the availability, patient, support and understanding shown. Sure that without your help and without your recommendations, my growth of learning in this area and the realization of this dissertation would not be possible.

In parallel, I want to thank my parents, not only for giving me education and values, but also for giving me all the necessary bases and giving me the opportunity to choose the path for my future. Thank you for all the love and proud you feel for me. It remains to thank also to my brother. For being an example of humility, dedication and persistence. Thank you for teaching me constantly to never give up and always fight for what makes me happy.

To my boyfriend. Thank you for playing the perfect role as a boyfriend, best friend and confidant. For being my unconditional support, by patience, for all the trust you put in me and for never doubting my abilities.

I have to thank my friends that have proved an important source of motivation and encouragement for me.

Lastly, I would like to thank the "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" project funded by the Northern Regional Operational Program (NORTE 2020) under the Portugal Partnership Agreement and through the European Regional Development Fund (ERDF) for the provision of data used to carry out this project.

Sofia Novais

"Para ser grande, sê inteiro: nada
Teu exagera ou exclui.
Sê todo em cada coisa. Põe quanto és
No mínimo que fazes.
Assim em cada lago a lua toda
Brilha, porque alta vive"

Ricardo Reis

Contents

List of Figures

List of Tables

Abbreviations and Symbols

ADReCS	ADverse Reaction Classification System
ADEs	Adverse Drug Events
ADR	Adverse Drug Reaction
ANN	Artificial Neural Networks
AUC	Area Under the [ROC] Curve
CBF	Content-Based Filtering
CF	Collaborative Filtering
ChEBI	Chemical Entities of Biological Interest
CLARA	Clustering LARge Applications
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CTCAE	Common Terminology Criteria for Adverse Events
CTD	Comparative Toxicogenomics Database
CV	Cross Validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DDI	Drug-Drug Interaction
DM	Data Mining
DO	Disease Ontology
DOGMA	Developing Ontology-Grounded Methods and Applications
EMA	European Medicines Agency
EU	European Union
FAERS	FDA Adverse Events Reporting System
FDA	Food and Drug Administration
KDD	Knowledge Discovery in Databases
LOOCV	Leave-One-Out Cross Validation
MedDRA	Medical Dictionary for Regulatory Activities
MF	Matrix Factorization
NCI	National Cancer Institute
NCBI	National Center for Biotechnology Information
NCSU	North Carolina State University
NIEHS	National Institute of Environmental Health Sciences
OAE	Ontology of Adverse Effects
OMS	Organização Mundial de Saúde
OWL	Web Ontology Language
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization
WHO-ART	World Health Organization-Adverse Reaction Terminology
PAM	Partitioning Around Medoids
PhV	PharmacoVigilance

PIDM	Programme for International Drug Monitoring
RAM	Reacção Adversa a Medicamentos
RBF	Radial Basis Function
RS	Recommender System
SAR	Structure-Activity Relationship
SE	Side Effect
SIDER	SIDe Effect Resource
SOM	Self-Organizing Map
SPC	Summary of Product Characteristics
SRS	Spontaneous Reporting Systems
SVM	Support Vector Machine
UMC	Uppsala Monitoring Centre
VAERS	Vaccine Adverse Events Reporting System
QSAR	Quantitative Structure-Activity Relationships
QSPR	Quantitative Structure-Property Relationships
THIN	The Health Improvement Network
XML	eXtensible Markup Language

Chapter 1

Introduction

This chapter presents a small contextualization of the theme and addresses the main objectives of this dissertation.

Finally, the structure of the report is presented.

1.1 Contextualization

Currently, Adverse Drug Reactions (ADRs) are a significant public health concern, not only for patients as well as for the health professionals, as they can potentiate serious injury and even lead to mortality of individuals. It is considered that these noxious and unintended effects are a significant reason for the current mortality. About 100,000 deaths are reported annually in the United States. It is the fourth cause of death. At the same time, the Adverse Drug Events (ADEs) potentiate the prolongation of hospitalization by about 3-6%, which translates into an increase in hospital costs that can reach millions of dollars [4, 5, 6].

Thus, continuous monitoring and prediction studies of drugs after marketing is essential, since the information gathered during the pre-marketing phase may be insufficient and incomplete. The use of some drugs may prove unsafe and pose risks for the patient due to them characteristics. It is necessary to consider medication errors such as the dose, duration of treatment or even medical malpractice, but even when the drugs are used rationally, undesirable effects may occur during treatment [7, 8, 9].

It is recognized that ADRs are a major global concern and therefore have been the subject of ongoing studies and research aimed primarily at increasing patient safety and protection. Computer technology has been playing a crucial and extremely important role in the health care. It is considered that the growth of health knowledge is directly proportional to the increase in the development of information technology. Prevention of errors is a primary reason for use of technology in pharmaceutical practice. Recently there has been a lot of research in developing to predict ADRs. The development of predictive procedures would help health professionals, since they could avoid many unwanted and unknown serious effects before marketing the drug. In parallel, these methods could also reduce hospital costs. Preserving the quality, safety and efficacy

of all drugs and ensuring the well-being of the population is the determining factor in this matter [10, 11, 12].

1.2 Objectives

The main goal of this dissertation is to assess how useful Data Mining methods and algorithms can predict the adverse effect of drugs. Thus, several key points have to be studied and addressed.

In the first stage, the main objectives are not only study and understand the basic concepts inherent to the adverse effects of drugs, but also analyze sites and databases where data and information concerning ADRs is available. After this research, it is important to investigate studies and methods that already exist to predict these effects and, finally, explore the tools of Data Mining and assimilate the interest of this method in predicting adverse effects of drugs.

The idea is to choose drugs, whose effects are known and available in a reliable database, and use a Data Mining tool to predict such known effects. That is, in order to verify if the tool used is feasible to predict unknown adverse effects of drugs, is performed a simulation where if supposes that adverse effects of selected drugs are not known and, through the crossing of data of other drugs, it is intended discover the adverse effects of drugs.

Thus, it is extremely important to put two research questions:

H1) can data mining software be useful for the process of prediction and explanation of adverse drug effects?

H2) does the use of additional information from the Chemistry domain improves the performance of data mining algorithms?

The answer to these questions is debated in the Chapter 5, in section 5.1.

1.3 Structure of the Report

This report has five chapters. The current chapter has presented the context of the dissertation, its motivation and its objectives. Chapter 2 introduces the state-of-art of adverse effects of drugs and Data Mining concepts. Various tasks and tools available in Data Mining are also described in Chapter 2. This chapter also surveys relevant ontologies and standards. Chapter 3 focuses primarily on project implementation. It presents not only a description of the data used and its pre-processing, but also details the processes performed and the algorithms used. The results obtained and some conclusions of these are addressed in Chapter 4. Finally, Chapter 5 describes the most relevant conclusions and suggests future work.

Chapter 2

Survey on Adverse Drug Reactions and Data Mining

This chapter surveys the essential topics of the domain of this study. It approaches the basic concepts related to drugs and adverse drug reactions and their underlying factors. The risks involving the use of drugs are reported together with the importance of monitoring of drugs in the post-marketing phase. Finally, it is presented a description of Data Mining state-of-art methods and algorithms and of the tools that currently can be used to predict adverse drug reactions.

2.1 Adverse Drug Reactions

As is general knowledge, the use of drugs [1] is very common, most often, prescribed by competent authorities as doctors or health professionals, in order to treat diseases or improve a health problem. Although drugs are considered generally effective in the treatment for the disease they were prescribed, the impact of their use in society has several perspectives. Drugs can affect the body of those who use them in unexpected and harmful ways for a variety of reasons, which may influence their medicinal and therapeutic efficacy [2], [3], [7, 13, 14].

All commercially available and marketable drugs have previously been subject to a number of pre-clinical tests, in order to prove their efficacy and safety. Thus, any drug is only approved for sale in the market if it is found that it has more benefits than risks for patients. However, because knows itself that no drug is risk free, surveillance of drugs after marketing is essential and extremely necessary [9, 14, 15].

[1] A medicinal product is any substance or combination of substances presented as possessing curative or preventive properties for diseases in humans or their symptoms or that may be used or administered to humans in order to establish a medical diagnosis or, by exerting a pharmacological action, immune or metabolic, to restore, correct or modify physiological functions.

[2] "Interações medicamentosas." Available in http://www.ebah.pt/content/ABAAAAjBUAE/interacoes-medicamentosas, accessed last time in 22-10-2016

[3] "Medicamento." Available in http://www.apifarma.pt/apifarma/areas/saudehumana/Paginas/default.aspx, accessed last time in 23-10-2016

The World Health Organization (WHO) defines Adverse Drug Reaction (ADR) as "any harmful or undesired effect which manifests itself after drug administration at doses normally used in man for the prophylaxis, diagnosis or treatment of a disease" [7, 8, 9].

There are certain factors that may be at the origin of possible occurrences of ADRs. This concludes that there is a higher incidence of ADRs in certain situations. It is then observed that the age and gender of the patient, the number of drugs administered and the interaction that occurs between them, the pathological condition, the existence of allergies and genetic factors could promote the occurrence of an ADR [14].

Adverse Drug Reactions classification

A lot of research has been done recently on ADRs. It is currently possible to classify them according to their severity and causality. In 1999, Wills and Brown [15] proposed an organization for to classify the ADRs, dividing them into nine main categories:

Reaction type A - usually depend on the dose, can be predictable, have low mortality and occur more often. Eg., poisoning digoxins in patients with kidney disease;

Reaction type B - for the most part are pharmacologically predictable, involves the interaction with the host microorganism and disappear with the removal of the causative agent. That differs from type A by the fact that it occur in the organism physiology. Eg., superinfections with the use of broad-spectrum antibiotics;

Reaction type C - concerns the concentration of the pathogenic agent and chemical characteristics and not the pharmacological effect of the drug. Eg., gastrointestinal injury by local irritant;

Reaction type D - depends on the physical nature of the preparation (formulation). The reaction ends when the drug is withdrawed or replaced. Eg., inflammation around implants;

Reaction type E - derived of the withdrawal or reduction in dose of drug and is pharmacologically predictable. Eg., tricyclic antidepressants;

Reaction type F - appears only in susceptible individuals, genetically determined. Eg., hemolysis with the use of sulfonamides for disabled glucose 6-phosphate dehydrogenase;

Reaction type G - genotoxic effects of medicines cause irreversible genetic damage. Eg., thalidomide;

Reaction type H - pharmacologically unpredictable, it is not dose-related, arises from the activation of the immune system and only disappears with drug withdrawal. Eg., anaphylactic shock caused by penicillin;

Reaction type U - despite not knowing the mechanisms of action, it is known that does not fit the above categories. Eg., vomiting after anesthesia.

Relatively to severity, ADRs are categorized into: [7, 9, 16].

Light - does not involve the withdrawal of the drug or require additional treatment;

Moderate - does not involve removal of the drug, but is necessary a modification of treatment;

Serious - involves drug withdrawal and cessation of treatment, and can be fatal or result in persistent or significant disability and requires hospital care;

Lethal - fatal to the patient.

Similarly, Karch and Lasagna in 1975 [4] proposed the following ADRs classification according to causality: [7, 9, 13].

Defined - meets a chronological sequence and follows a known response;

Probable - meets a chronological sequence and follows a known response, however cannot be explained by the known characteristics of the patient's clinical status;

Possible - meets a chronological sequence and follow a known response, however it may have been a result of the clinical state of the patient or other treatments that are being performed simultaneously;

Conditional - meets a chronological sequence, but does not follow a known response and cannot be explained by the known characteristics;

Doubtful - any reaction that does not follow the rules described above.

Patient-related factors that predispose to Adverse Drug Reactions

There is great disparity in the way people respond and interact with the drugs. Thus, although all tests and clinical tests performed in order to ensure not only in produce satisfactory results such as the improvement in the quality of life of the population, they have some limitations that must be taken into consideration. Some tests in animals are insufficient to provide safety to human patients. The number of patients involved in these studies is very limited. The short duration of tests and the amounts of the doses, which may not correspond to the time/dosage actual required to treatment. The exclusion of individuals who have diseases or are old/children. All of these are considered the main conditions limiting these clinical trials [5], [6], [4, 14, 17].

It is extremely important to note that there are certain groups of individuals more susceptible to the occurrence of ADRs, such as the elderly/children/women or individuals who have a clinical

[4] "Trabalho de Farmacologia sobre reações adversas", Marcelo Santos. Available in http://www.ebah.pt/content/ABAAABIYwAB/trabalho-farmacologia-sobre-reacoes-adversas?part=2, accessed last time in 16-01-2017

[5] "Farmacovigilância: Reações adversas a medicamentos", August 2008. Available in http://www.ufrgs.br/boletimcimrs, accessed last time in 22-10-2016

[6] What are adverse drug reactions? Available in http://www.adr-database.com/What%20are%20ADRs.html, accessed last time in 23-10-2016

history of diseases. It is also verified that there is a direct relation between the number of drugs that a patient takes, that is, the probability of an ADR occurring increases with the number of drugs ingested by the patient. The administration of large doses of drugs promotes the occurrence of ADRs. It is also important to note that individuals with allergies or those with liver, heart or kidney disease are more easily affected. For this reason, this type of population will have to have adequate and continuous clinical monitoring [14, 16, 18].

Drug-Drug Interaction

Several previous studies claim that a very relevant factor that can increase the number of ADRs episodes is the simultaneous consumption of more than one drug, since drugs are chemical compounds that can interact with each other, triggering, in this way unwanted responses and sometimes having toxic effects. Thus, it is essential to study not only the occurrence of ADR episodes in isolation, but also the possible drug interactions that may exist and the relationships that can be created between the two [16, 18].

Currently, it is found that the majority of the medicated population consumes more than one drug. On the one hand, we have situations in which the taking of these is independent, that is, they are used for different purposes of treatment (for example, one drug serves to lower cholesterol and another drug serves to raise blood pressure) and, on the other hand, the taking of these may be dependent, that is, when they are used together for the same treatment. The simultaneous taking of several drugs is one of the most worrying and relevant issues in clinical practice, and health professionals should be fully aware of the patients' medical history: drugs taken, doses administered, period of time taken, diseases, etc., in order to guarantee their safety.

As is well known, the prescription of drugs is totally performed with the purpose of guaranteeing benefit to the patient, contributing to their well-being and significantly improving their health and quality of life, thus, the simultaneous use of medication is a therapeutic strategy widely used and, most of the times, necessary, which enhances the effectiveness of the drugs and, consequently, the beneficial effects of the treatment. At the same time, however, their interaction may not only lead to unwanted responses, but may also change the intended effects, since they are chemical substances that may be interacting with each other, with the receptors or with agents around them. The chemical responsible for the therapeutic action is called the active principle of the drug. The activity of the drug can be determined by its chemical structure. Thus, after it comes into contact with the body or after it interacts with another drug, a change in the molecular structure of the drug's active principle may occur, and thus, the occurrence of ADRs will be induced [2], [7], [8], [18, 19].

Briefly, the Drug-Drug Interaction (DDI) comes from the simultaneous ingestion of two or more drugs and consists in the alteration of the effects of the individual drugs, which may result either in the increase or decrease in toxicity or most often in the increase/reduction in terms of

[7] "Farmacologia molecular." Available in http://antonini.med.br/geral/farmol.html, accessed last time in 17-11-2016

[8] "Divisões e conceitos da farmacologia", July2015. Available inhttp://luizcarlosfarmaceutico.webnode.com/news/divisoes-de-conceitos-de-farmacologia/, accessed last time in 17-11-2016

therapeutic efficacy, DDIs are dangerous and endanger the lives of patients. Several studies have demonstrated the direct relationship between the number of drugs consumed and the incidence of ADRs. It is observed that the probability of a patient having an increases with the number of drugs taken (ranging from 3-5% for patients who take 2 to 3 drugs, increasing to 20% or more, in patients consuming 10 to 20 drugs). A recent study in a set of hospitals in the USA revealed that about 22% of the ADRs are due to DDIs. This value reveals the extreme importance and necessity of methods and tools that help and guide health professionals so that they can make credible decisions with a lower risk of toxicity to patients. It is also notorious that the incidence of DDI is more susceptible both in the elderly and in the patients who are hospitalized, since they have a weakened organism, as well as in children, since they may not yet have the organism fully prepared. In this way, the role of health professionals is extremely important and has a great responsibility to combat this problem, since taking into account the factors mentioned above and with knowledge of the characteristics of the drugs, many DDIs could be avoided [20, 21, 22].

Regarding the nature, DDI may be **pharmaceutical**, **pharmacokinetic** or **pharmacodynamic**. For D.G. Grahame-Smith and J. K. Aronson [7], the **pharmaceutical** DDI, also called drug incompatibility, comprises "physical-chemical events that result in the loss of activity of one or both of the drugs." These occur before the administration of the drugs, that is, outside the body, when the drugs are placed and mixed in the same container. Usually, this type of DDI can be observed in the preparation of drugs, since they usually result in color/consistency or precipitation changes and may cause the formation of a new compound or decrease/inactivate the desired activity of the drugs. It is known that the drugs most susceptible to this type of DDI are those that require continuous infusion (Example: Penicillin is capable of inactivating the aminoglycosides when they are mixed). On the other hand, it is reported that **pharmacokinetic-type** DDIs are the most frequent ones, and are unfortunately the most difficult to predict, since they appear in drugs with unrelated active principles. This type of DDI comprises significant changes in parameters such as plasma concentrations and affects mechanisms such as absorption, distribution, metabolization or excretion of the drug. Absorption of the medicinal products may be impaired as they may become insoluble in the body due to, for example, gastrointestinal pH changes that may occur when the drugs interact. Several studies mention that the antacids drug group has the ability to decrease the absorption of other drugs. The distribution process is characterized by the arrangement of the drug in the tissues, and for this, many drugs bind to plasma proteins. Thus, it is easy to see that when there is simultaneous taking of more than one drug, there may be competition for these plasma proteins, and the connection with them will occur with the drug that has the most affinity to it, thus avoiding the correct distribution of drugs. This happens when warfarin and coral hydrate are ingested in parallel. In this case, coral hydrate is responsible for non-binding of warfarin to plasma proteins, thereby increasing the plasma concentration of warfarin. During the metabolism the transformation of drugs, through biochemical reactions catalyzed by enzymes essentially hepatic occurs. Presently, some drugs are known that can both induce or inhibit enzyme activity, which may cause the loss/reduction of the action of the drug to be ingested in concomitance. Fentoin, defined as an enzyme inducer, when interacting with oral contraceptives cuts the effect of them.

Excretion, that is, elimination of drugs happens essentially in the kidneys. The rate of this can be altered, that is, it may take more or less time than expected, by the interaction of drugs, causing the increase or decrease of concentration of the drug in the body, respectively. It is known that the time of action of the antibiotic penicillin is amplified when it is taken together with the probenecid drug, since the second has the ability to inhibit the secretion of the former, which can produce a toxic effect. Finally, it is also important to approach the DDI of the **pharmacodynamic** category, since it is responsible for interfering in both the biochemical and physiological effects of the drug, thus occurring at the site of its action, at the level of the pharmacological receptors. Depending on the type of effect it produces, the response of this type of DDI can be of two types. If the product is potentialization of the effect, that is, if the effect of the drug is similar to that predicted, but increased, synergism occurs. This happens when the drug acts at different pharmacological receptors. If, on the other hand, the result is a worsening of side effects, that is, if the effect of the drug is the opposite of what is intended, there is antagonism. This occurs when drugs compete for the same pharmacological receptors, being that one of which antagonizes or suppresses the other. Potentialization of the effect arises when the combined effect of two or more drugs is greater than the isolated effects of each drug. Usually, these effects are predictable and can be used to benefit and maximize therapeutic action, however, they can also cause toxicity. An example of DDI of this type is the association of medications such as aminoglycoside and vancomycin, in which the production of toxic effects to the patient occurs. Analogously, the combination of diuretics harms the effects of the drug digoxin and is therefore an example of an IDD of an antagonistic pharmacodynamic nature [9], [10], [18, 20, 21].

From another perspective, DDIs can be classified as **mild**, **moderate** or **severe**. They are called **mild**, when it is not necessary to change or withdraw the drugs in question. For example: the effect of the diuretic furosemide is decreased by the analgesic acetaminophen, however their combination does not result in toxic effects for the patient. When the patient's clinical condition worsens, DDI is said to be **moderate**. For example: Simultaneous use of the antibiotic rifampicin and the anti-tuberculosis drug called isoniazid. In case of irreversible/permanent effects and possibility of death of the patient, the DDI is designated as **severe**. For example: The combination of the antihistamine terfenadine and the antifungal ketoconazole potentiates the occurrence of cardiac arrhythmias. There is some relevant information about drugs that should be taken into consideration at the time of prescription, in order to avoid certain DDIs. Thus, drugs such as aspirin, warfarin and enzyme inhibitors are known to have the potential to interact with other drugs and potentiate toxic effects [2], [7].

In conclusion: despite all the analyzes performed, the existing information about drugs, the previously reported factors that are likely to induce the occurrence of an DDI and the enormous knowledge of health professionals, it is still difficult to predict with absolute certainty a DDI.

[9] "Interação entre medicamentos: mantenha-se alerta." Available in https://advancecare.pt/interacao-entre-medicamentos-mantenha-se-alerta/,accessed last time in 23-10-2016

[10] "Principais interações medicamentosas e drogas associadas com efeito", 2008. Available in http://boaspraticasfarmaceuticas.blogspot.pt/2008/10/principais-interaes-meicamentosas-e.html,accessed last time in 23-10-2016

This occur because each patient is different and can react differently to drugs. For this reason, the role of health professionals is very important, as they should analyze each case in depth and pay special attention to the symptoms that arise from DDIs. It is certainly essential to have other options. Thus, it is necessary to select substitute drugs and, if this is not possible, it is essential to adjust doses, implement alternative therapeutic strategies and regularly monitor the patient, in order to try to ensure maximum patient safety and treatment efficacy. The health professionals should also notify the competent authorities of the occurrence of an ADR in order to alert, protect and avoid other similar occurrences [2], [9], [11], [12], [21].

Post-marketing Safety monitoring of drugs

Continuous surveillance of medicinal products after they have been marketed is extremely necessary and essential, since not only there are the numerous factors mentioned previously (mentioned in 2.1) which make the tests carried out at the pre-marketing stage incomplete and imperfect, especially with regard to possible adverse reactions, as well as changes in the risk/benefit ratio may occur over time as new information and occurrences on the use of such a medicinal product arise. Studies have estimated that in about half of the drugs new adverse reactions are only known after being approved for sale. Thus, after the drugs are available on the market, care should be taken to identify any new risks as quickly as possible and thus to ensure maximum safety and protection in order to avoid harmful harm to patients [5],[13], [4, 12, 23].

The awareness about the need to implement drug surveillance systems whose sale was already allowed came only in the 1960s when the tragic accident with the drug thalidomide prescribed for pregnant women, responsible for the birth of infants with congenital malformations, occurred. This event forced the withdrawal of the drug and since then several cases have been known whose drugs were withdrawn from the market because they put the patients' life in danger. It is important to note that, despite the enormous concern regarding global health that this implies, pharmaceutical companies are also affected by billions of euros [4, 11, 24]. Since then, several countries began to use **PharmacoVigilance (PhV)** systems, that is, drug safety monitoring systems, to identify and detect ADRs, improve public health and avoid financial risks. PharmacoVigilance is defined by WHO [25] as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" [21], [10, 26, 27]. The aim of this system is to recognize unknown DDIs, to detect ADRs early on, to improve the rational use of drugs and, consequently, to reduce the mortality rates that come from ADRs [10, 12, 15].

Currently, the most used tool and that constitutes the primordial source in PhV is denominated spontaneous notification. This is the communication of suspected ADRs, signs or symptoms detected by both health professionals and patients, in order to inform and alert the regulatory and responsible entities of the drugs so that they analyze the data provided and verify the veracity of

[11] Clinically important, common drug-drug interactions (ddis). Available in https://crediblemeds.org/healthcare-providers/drug-drug-interaction, accessed last time in 22-10-2016

[12] "Novidades sobre a indústria farmacêutica no brasil." Available in http://www.hipolabor.com.br/blog/2015/05/04/hipolabor-explica-voce-sabe-o-que-e-interacao-medicamentosa/,accessed last time in 22-10-2016

[13] Pharmacovigilance, 2015. Available in http://www.ufrgs.br/boletimcimrs, accessed last time in 22-10-2016

the information, in order to continue to ensure the quality, efficacy and safety of the medicinal product and, if necessary, to suspend or withdraw it from the market. These participations are conducted in Spontaneous Reporting Systems (SRS) at national level and are stored in databases and sent to the WHO Collaborating Centre for International Drug Monitoring (Uppsala Monitoring Centre - UMC), so that information is shared between the countries and other European Medicines Agency (EMA) [14], [23, 25, 28]. Although this method contributes essentially to the continuous protection of the overall health and improvement of therapeutic practice and is very relevant in post-marketing surveillance, it presents a great disadvantage and limitation, since it does not allow a direct benefit to the patient, that is, this system is only applicable after the patients have been targeted by the ADRs [15, 24, 29].

Analogously, a number of studies and research have emerged whose main purpose is the development of systems and the use of computational methods, such as Data Mining techniques. Predictive methods may predict possible ADRs before they actually occur, significantly improving public health and safety in drug taking [4, 24, 26].

Web Data Repositories

With this "collective" participation is possible to alert the competent entities about occurrences of ADR. To collect information about ADRs and to compare experiences it requires the participation and collaboration of health professionals, regulators and users [10, 11, 12].

By 1968, WHO formed the **Programme for International Drug Monitoring (PIDM)** to gather information on ADRs at the international level. The main objective of this program is to facilitate the communication between countries to facilitate the identification of ADRs through reports and case reports. WHO coordinates the system for quantification and detection of adverse reactions by the Collaborating Center of Uppsala Monitoring Center, whose main function is to organize the participation of all 120 countries that maintain VigiBase [15], a worldwide database on ADRs. Typically, these countries provide similar reporting forms in order to establish a standard with a certain level of consistency in the information collected. This database provide public access to nearly 10 million ADRs. To do this, it is only necessary to enter the name of the drug at the research site. This research makes it possible to see the types of reactions that were reported for the product, broken down by category and frequency, geographic distribution, age group, patient gender and number of ADRs per year. The data contained in VigiAcess (http://www.vigiaccess.org/) are aggregated so that both the entity of the notifying person and the country where the notification was made be kept confidential [16], [17], [12, 14].

[14] Report side effects. Available in http://www.medicinesauthority.gov.mt/adversedrugreactions?l=1, accessed last time in 26-10-2016

[15] VigiBase. Available in http://www.umc-products.com/DynPage.aspx?id=73590mn1=1107mn2=1132, accessed last time in 19-01-2017

[16] Who makes drug side effects easier to find with new database, April. Available in http://www.raps.org/Regulatory-Focus/News/2015/04/23/22034/WHO-Makes-Drug-Side-Effects-Easier-to-Find-With-New-Database/, accessed last time in 26-10-2016

[17] VigiAccess. Available in http://www.vigiaccess.org/, accessed last time in 18-11-2016

There are several databases available on the Internet for ADR reporting. Most are free and with public access to the reports provided and to both the data and the conditions under which these ADRs were experienced.

All reports concerning ADRs within the European Union (EU) as well as reports from outside the EU submitted by the marketing authorization holders in accordance with EU law are maintained by the EMA in a database called **EudraVigilance** [18]. In order to ensure the anonymity of data subjects and the protection of data, these are not available in full to the general public. Aggregate data can be seen by age group, gender, type of suspected side effect and outcome [19], [25].

DrugBank is a freely available database storing a large amount of molecules information. It is considered a unique biology and chemistry resource that contains detailed data not only of drugs, but also of the amino acid sequence, structure and route of administration. It has approximately 8261 drug entries. It allows one to search by medication, category, reaction, pathway, gene, target protein, class or indication and collects/reviews information in more than 50 databases/web applications. Data are available in XML format [20]. For our study it is important that DrugBank has information concerning drug-drug interaction. It is also valuable for classification tasks as we have in the study the hierarchical taxonomy of the molecules.

There is also a database called **ADverse Reaction Classification System (ADReCS)** which provides the standardization and hierarchical classification in four levels of ADR terms. The information contained in this database originates from medical repositories such as DailyMed (http://dailymed.nlm.nih.gov/dailymed/about.cfm), MedDRA (http://www.meddra.org/), DrugBank (http://www.drugbank.ca/) or PubChem (https://pubchem.ncbi.nlm.nih.gov/). It contain information on 1698 drugs and 6972 ADR terms, in a total of 157246 drug-adverse events. Academic users can access this data free of charge at http://bioinf.xmu.edu.cn/ADReCS/download.jsp. Data are available in XML format [21], [5].

OpenFDA was designed in 2014 with the main objective of allow public upload of ADR cases. Is an excellent and complete data source. It contains and provide numerous personal data, such as the list of drugs administered by them and their associated ADRs. Offers the possibility for users to download structured data, namely adverse events and drug labels. These informations are very important and useful since it can allow and facilitate the study of possible DDIs [22].

The well-known regulatory agency of drugs, US Food and Drug Administration (FDA), is responsible for, among other things, ensuring the safety and efficacy of drugs in order to preserve the health of the population. With this objective in mind, FDA has developed the **FDA Adverse Event**

[18] European database of suspected adverse drug reaction reports. Available in http://www.adrreports.eu/en/eudravigilance.html, accessed last time in 19-01-2017

[19] Medicines authority. Available in http://medicinesauthority.gov.mt/missionobjectives?l=1, accessed last time in 19-11-2016

[20] DrugBank. Available in https://www.drugbank.ca/about, accessed last time in 09-06-2017

[21] Xiamen University. Adrecs, 2014. Available in http://bioinf.xmu.edu.cn/ADReCS/, accessed last time in 19-11-2016

[22] About openFDA. Available in https://open.fda.gov/about/, accessed last time in 18-01-2017

Reporting System (FAERS) [23] that contains adverse reactions and medication error reports. It has a program called MedWatch, enabling the report of possible ADRs. At the moment (2016), it has already nearly 2 million reports. It is important to note that since these are spontaneous reports made by people, mostly without any medical knowledge, the data contained in the reports is not completely reliable [24], [4].

SIDe Effect Resource (SIDER) [25], hosted by the European Molecular Biology Laboratory, provides access to information on marketed drugs and stores around 5880 data on ADRs. The information contained in this database is standardized using the Medical Dictionary for Regulatory Activities (MedDRA) and comes from both public and bulletin documents. This data can be downloaded in text format. In addition to allow users to search for drugs and ADRs, SIDER features a GitHub repository where they can report error detection in the information made available [11, 30].

The partnership of two companies - INPS (IN Practice Systems) and EPIC - gave rise in 2003 to the database of the **The Health Improvement Network (THIN)** [26] in the United Kingdom. This database covers the electronic medical records of 11.1 million patients from hospital practice that were processed, validated and kept anonymous by CSD Medical Research UK. THIN contains two types of data called "therapy" and data defined as "doctors". The first relates essentially to the details of patient prescriptions. The second stores both symptoms, diagnoses and medical interventions. The diagnoses provided are coded using hierarchical codes, in order to follow a pattern [27].

There are also several online forums such as PatientsLikeMe [27] or blogs where it is possible to share patient self-reports about one's experiences with the use of medication. The reliability of the information provided is questionable, but this sharing of information is always relevant and necessary, since it can provide useful information on some drugs [10, 26].

In order to understand how adverse reactions occur, it is necessary to perceive the mechanisms they take on drugs at the molecular level. Both access to chemical and biological information may aid in understanding the molecular mechanisms involved and can be found at the source of adverse reactions. There are currently some servers that collect this information. The **IntSide web** server is updated each 6 months and contains about 1175 side-effects of 996 drugs. These are divided, based on an analysis performed by Miquel Duran-Frigola and Patrick Aloy [31], into eight categories according to biology and chemistry: Targets, Proteins, Pathways, Processes, Functions, Fragments, Scaffolds and Structural Terms. Users can scan up to three Side Effects (SEs) at a time and the server will produce different networks displaying the common biological

[23] FDA Adverse Event Reporting System. Available in https://open.fda.gov/data/faers/, accessed last time in 19-01-2017

[24] U.S FOOD DRUG ADMINISTRATION. Available in http://www.fda.gov/default.htm, accessed last time in 18-11-2016

[25] SIDER 4.1: Side Effect Resource. Available in http://sideeffects.embl.de/, accessed last time in 19-11-2016

[26] The health improvement network (thin) research team, April 2015. Available in https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database, accessed last time in 19-11-2016

[27] Available in https://www.patientslikeme.com/, accessed last time in 19-11-2016

and chemical characteristics associated with each SE. These networks have the identification of complex mechanisms [28].

Comparative Toxicogenomics Database (CTD) was developed at the North Carolina State University (NCSU) and is publicly available for free via TOXNET [29], supported by funds from the National Institute of Environmental Health Sciences (NIEHS). This database gathers manually curved information on the interactions that occur between molecular mechanisms: chemical-gene/protein interactions, chemical-disease interactions and gene-disease interactions. The main objective is to provide an insight into how chemicals relate to and influence human health, helping to develop relationships that may be at the root of side effects. The gene for cross-linked CTDs (symbols, names and data) is derived from the Gene Database of the National Center for Biotechnology Information (NCBI), a division of the US National Library of Medicine [30], [31], [32].

Molecular Descriptors and Tools

Several studies have shown that the Structure-Activity Relationship (SAR) of substances can be investigated using molecular descriptors. For R. Todeschini and V. Consonni [32], "the molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment". Normally these are used for the purpose of predicting the properties of substances, manipulating and analyzing chemical structural information in order to classify chemical structures and seek a similarity between them [33],[34], [35], [33].

According to Katritzk and Gordeeva [34], molecular descriptors are divided into: **topological**, **electronic**, **geometric** or **combined**. The so-called **topological** descriptors are obtained by the two-dimensional (2D) graphical representation of molecules and characterize both size and molecular form. In turn, those that are calculated through quantum chemistry are known as **electronic** descriptors. They allow the expression of both the electronic properties of the molecules, as well as their interactions. M. Karelson and V. S. Lobanov [35], in 1996, verified the importance and usefulness of this type of descriptors in the development of Quantitative Structure-Activity Relationships (QSAR)/Quantitative Structure-Property Relationships (QSPR) with the chemical, physical, biochemical and pharmacological properties of the compounds. The differentiation of molecules by their topology is a function of the **geometric** descriptors. The use of this type of

[28] Intside, 2015. Available in http://intside.irbbarcelona.org/index.php, accessed last time in 24-11-2016

[29] Available in http://toxnet.nlm.nih.gov, accessed last time in 24-11-2016

[30] Comparative Toxicogenomics Database. Available in http://ctdbase.org/, accessed last time in 24-11-2016

[31] Fact Sheet Comparative Toxicogenomics Database (CTD), 2014. Available in https://www.nlm.nih.gov/pubs/factsheets/ctdfs.html, accessed last time in 24-11-2016

[32] What is a molecular descriptor, 2000. Available in http://www.moleculardescriptors.eu/tutorials/T1_moleculardescriptors_what_is.pdf, accessed last time in 24-11-2016

[33] An integrated web-based platform for molecular descriptor and fingerprint computation, March 2015. Available in http://www.scbdd.com/chemdes/, accessed last time in 17-11-2016

[34] Molecular descriptors: An introduction, 2006. Available in http://infochim.u-strasbg.fr/CS3/program/material/Todeschini.pdf, accessed last time in 24-11-2016

[35] "Descritores moleculares para aprendizagem automática ("Machine learning")." Available in http://joao.airesdesousa.com/agregacao/slides_2013/descritores_QSPR_slides.pdf, accessed last time in 24-11-2016

descriptors requires knowledge of the three-dimensional (3D) coordinates of the atoms of the molecule. Finally, the descriptors are called **combined** when they are calculated by computer-assisted methods. This category of descriptors is responsible for grouping information about the topological environment of each atom and about the electronic interactions due to the presence of other atoms in the molecule [34], [35], [36], [36].

The interconversion of molecular structures between representations constitutes a huge problem in computational modeling. Thus, tools are needed to facilitate the extraction and interpretation of chemical data [37].

Open Babel is an open source tool freely available from the Open Babel website and addresses the languages of chemical data. It is used for a wide variety of purposes, among which are the filtering, conversion, analysis and storage of molecular modeling data. Thus, this tool is a good option for the proliferation of multiple formats of chemical files, since it has the capacity to interconvert more than 110 chemical file formats [37], [37].

Fingerprints: The Open Babel tool has the ability to identify all the linear and ring substructures of a molecule. These are mapped to a bit sequence, known as the fingerprint of the molecule, of length 1024 using a hash function. In addition to having the ability to reduce the search time of a molecule, fingerprints also facilitate the identification of similar molecules.

SMILES - Represents the language used to identify a molecule. Such language complies with a set of rules. This allows the identification of a molecule by one and only one SMILES. This translates into an advantage in eliminating duplicates.

2D and 3D coordinates - Open Babel produces the 2D and 3D coordinates of a molecule through its SMILES identifier.

In Figure 2.1, it can be seen a representation of the three different ways of identifying the same molecule: SMILES identifier - considered a 0D structure (above); 2D (left side) and 3D (right side). These three structures can be identified and converted to one another using Open Babel.

Like as Open Babel, **PaDEL** is an open source tool used to calculate molecular descriptors and fingerprints. Molecular fingerprints designate the property of a molecule. Usually, they are presented in the form of bits or as vector representations to indicate, respectively, the existence of certain properties and the frequency of these. Currently, this software calculates approximately 1875 - molecular descriptors: 1444 are 1D and 2D descriptors and 431 are 3D descriptors - and about 12 types of fingerprints [38], [39], [38].

[36] What is a molecular descriptor, 2000. Available in http://www.moleculardescriptors.eu/tutorials/T1_moleculardescriptors_what_is.pdf, accessed last time in 24-11-2016

[37] Open babel: The open source chemistry toolbox, 2011. Available in http://openbabel.org/wiki/Main_Page, accessed last time in 17-11-2016

[38] PaDEL-Descriptor, July 2014. Available in http://www.yapcwsoft.com/dd/padeldescriptor/, accessed last time in 17-11-2016

[39] Chemical descriptors calculation. Available in http://crdd.osdd.net/descriptors.php, accessed last time in 17-11-2016

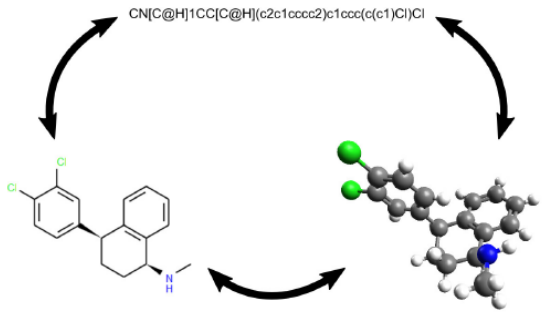

Figure 2.1: Identifiers of the molecule Sertraline

2.2 Data Mining

The amount of existing information and data stored in databases or files, in several areas, is quite large. To be able to take advantage of such data, powerful tools are required. Being able to analyze, interpret data and extract the most potentially useful knowledge from them, would enable use to make more conscious and relevant decisions [39, 40, 41].

Knowledge Discovery in Databases (KDD) is a technological answer to this need. Fayyad et al. [42, 43] describe KDD as being "a non-trivial process of identifying new valid, useful and understandable patterns". Thus, as the name implies, KDD aims at finding useful knowledge by extracting information from a database. It presents steps such as data selection, data cleansing, data transformation, being that according to Fayyad et al. [42, 43], "**Data Mining (DM)** is considered the main phase of this whole process and consists in the accomplishment of data analysis and the application of discovery algorithms that, under certain computational limitations, produce a set of patterns of certain data" [43, 44, 45]. In Figure 2.2, it can be seen a representation of the KDD process [40].

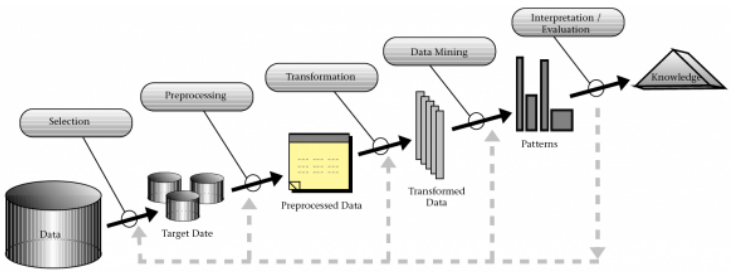

Figure 2.2: KDD Process.

DM is currently one of the most effective techniques for extracting knowledge from large

[40] Knowledge Discovery in Databases (KDD). Available in http://www.infovis-wiki.net/index.php?title=Knowledge_Discovery_in_Databases_(KDD), accessed last time in 17-11-2016

volumes of data. This exploration tool involves statistical methods, artificial intelligence techniques, namely machine learning, and specific algorithms and allows data segmentation, pattern/model discovery and significant correlations, and the extraction of structures [46, 47, 48]. Defined by Cabena et al. [43] as "a multidisciplinary field that combines knowledge machine techniques, pattern recognition, statistics, database and visualization, to extract information from large databases", DM has shown to be very promising in several areas, having satisfactory applicability in marketing, medicine, commerce, telecommunications, astronomy, fraud detection, among others [44, 49, 50].

In order to better answer the question in cause and to efficiently extract knowledge from the databases, different algorithms and tasks are applied, depending on the type of problem to be solved [39, 44, 49].

CRISP-DM

DM is characterized by several processes. Currently, the **CRISP-DM** (CRoss-Industry Standard Process for Data Mining) [51] is considered the most accepted standard. As can be seen in Figure 2.3, the CRISP-DM process is a six-stage cycle: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. These do not occur with unidirectional flow [43, 47, 50].

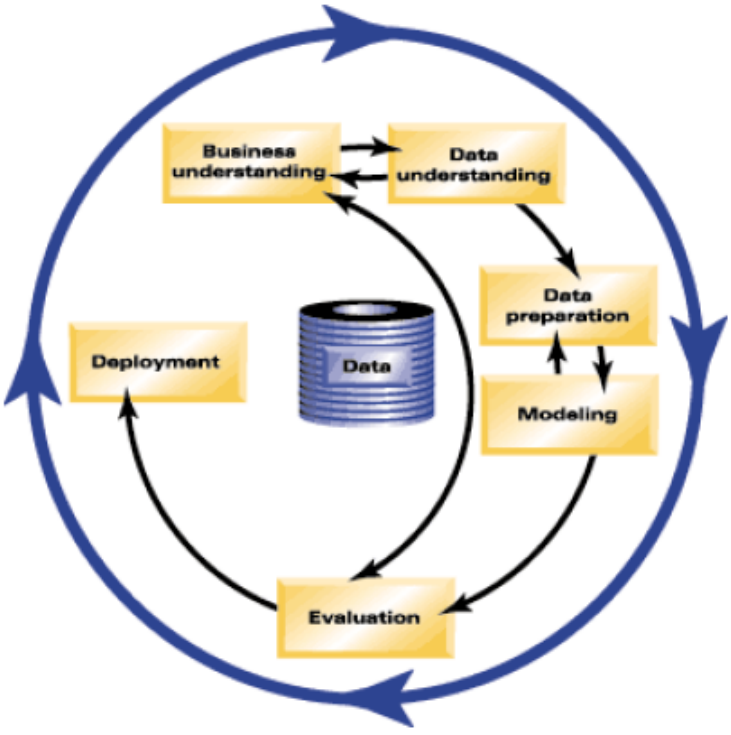

Figure 2.3: Phases of CRISP-DM [41].

1. **Business Understanding:** This phase is essentially characterized by the understanding of the objectives, from a business perspective, that are intended to be achieved through the use of DM. At the same time, a plan is implemented to achieve these objectives.

2. **Data Understanding:** It is necessary to know and identify the essential data of the problem, so that it can be clearly described. The data comes from a variety of sources and may be available in a variety of formats. In this step, one can use grouping techniques.

3. **Data Preparation:** In this step the data is prepared so that the DM methods can be used. To do this, the data cleaning process is used: filter, combine and fill empty values, so that the data will have the characteristics and quality required for the application of the DM methods to be performed and the results satisfactory.

4. **Modeling:** This phase involves the selection and subsequent application of DM algorithms. This choice is based on the objectives previously defined. The construction of a model is aimed at stability, that is, it is intended to make forecasts that are not only useful in decision making, but also remain true when applied to data not yet seen.

5. **Evaluation:** This step is extremely important, since it precedes the implementation of the model. Thus, it is necessary to evaluate in detail the model and to review what was done for its construction, in order to ensure that it responds adequately to the established objectives. To do this, we use specialists who will perform tests and validations, in order to guarantee the reliability and purpose of the model.

6. **Deployment:** Finally, it is necessary to organize and present knowledge in a way that the client can use. Deployment involves steps such as planning, monitoring and maintaining the plan, producing the final report, and reviewing the project. These, most of the time, are performed by the client and not by the analyst.

Data Mining Tasks

DM has the ability to perform tasks that can be categorized into two types: descriptive data mining (unsupervised learning) and predictive data mining (supervised learning). The first is characterized by presenting the data set and by highlighting the relevant general properties of the data, in order to discover patterns interpretable by the users. On the other hand, the second one is essential in the construction of models, being useful in the prediction of the behavior of new cases [40, 50, 52].

Commonly, a DM system can perform, among others, the following tasks: **Classification**, **Regression**, **Association Rules Discovery** and **Clustering**. The **Classification** is known as one of the most used predictive tasks, and can be supervised. Defined by Elmasri [42] as "the process of learning a function that maps (classifies) a given object of interest into one of the possible classes". The goal is therefore to assign an object, characterized by a set of attributes, to one of a finite set of labels (the class value). It involves a set of steps: data preprocessing, modeling, selection of learning/selection of resources and validation/evaluation. The model obtained by the

classification can be represented in the form of **classification rules (if-then)**, **Support Vector Machines (SVM)**, **Decision Tree**, **Random Forest**, **Naive Bayes**, among others [43, 53, 54].

Rule-based classification is often retrieved from a decision tree. Nowadays, Sequential Coverage algorithms such as AQ, CN2 and RIPPER, algorithms are used as a strategy for obtaining directly the rules [43, 53, 54].

Support Vector Machine - SVM is a supervised machine learning algorithm that is used both in classification and regression [42], [43]. This is characterized by being a binary non-probabilistic classifier. This is observed, since after taking an input dataset, it predicts, for each input taken, which of the two possible classes includes the input, separating a set of objects with different classes [44], [55].

Each training example is represented as a point in space. These are mapped so that a straight line divides them into categories. This line should be as far away as possible from the classes that are closest to the line. So, depending on which side of the line a new example is mapped, SVM predicts which of the classes this example belongs to. In some situations, the data are not linearly separated, and it is necessary for the SVM algorithm to resort to the kernel function. This is responsible for transforming the space of the training examples, so that they are presented in an n-dimensional space and that a hyperplane separates their classes [55].

Decision Trees are a very efficient and relevant algorithm type for predictive modeling machine learning [45], [46], [56]. The presentation of the data is done in the form of an inverted tree, that is to say, the root of the tree is in the top and this one is dividing in branches from the direction of upwards downwards. The input attributes are represented by the inner nodes of the tree. Each branch represents a possible result for this test. The classes that the algorithm is trying to predict are represented by the lower nodes, called sheet nodes [55].

The use of this algorithm is advantageous, since it is not only easy to interpret, it also does not require much preparation of the data and it can deal with missing values. In contrast, this algorithm can originate super-complex trees, causing overfitting [55].

In each recursion of any decision tree algorithm the following steps are used:

- Selection of an attribute A that will be divided;

- Subdivision of attribute A in disjoint subsets;

- A tree is returned with attribute A as root and with a set of branches on the lower level. Each branch will have a descending sub tree;

[42] 1.4. Support Vector Machines. Available in http://scikit-learn.org/stable/modules/svm.html, accessed last time in 27-05-2017

[43] Understanding Support Vector Machine algorithm from examples (along with code), by SUNIL RAY , 2015. Available in https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/, accessed last time in 27-05-2017

[44] Support Vector Machines (SVM) Introductory Overview. Available in http://www.statsoft.com/Textbook/Support-Vector-Machines, accessed last time in 27-05-2017

[45] 1.10. Decision Trees. Available in http://scikit-learn.org/stable/modules/tree.html, accessed last time in 27-05-2017

[46] Classification And Regression Trees for Machine Learning, by Jason Brownlee, 2016. Available in http://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/, accessed last time in 27-05-2017

- Generally, the tree stops branching when all examples have the same label value. In parallel, there are other conditions that cause the recursion to end: There are less than a number of instances or examples in the current sub tree; This can be adjusted through the parameter minimum size for split; No attribute achieves a given minimum gain relative to the existing tree before that division; This can be set via the minimal gain parameter; And finally, the maximum depth is reached. This can be adjusted through the maximal depth parameter.

There are several algorithms designed for the construction of decision trees: ID3, C4.5, CART and CHAID. These are essentially distinguished by the criterion that decides which next node to be explored and the type of test performed at each node inside the tree. In the set of experiences carried out in this dissertation (described in 3.1.2), the CART decision tree algorithm was used. In this type of algorithm, the choice of node A is made according to a selection criterion: metric Gini Index or Gini Impurity. Thus, the node that presents a smaller impurity measure is selected [55]. Such impurity can be calculated as follows:

$$GiniIndex(f) = \sum_{i=1}^{J} fi(1 - fi) \tag{2.1}$$

Being that:

- J is the number of classes;

- fi corresponds to the fraction of items labeled with class i.

Random Forest is considered a machine learning algorithm that generates a set of decision trees using sampling with replacement in the original dataset. The number of decision trees generated is specified by the programmer. This algorithm gives rise to a model that results from the voting among all the generated random trees. This means that, for example, a certain attribute "a" belongs to class "1", if this attribute was predicted by most trees generated as belonging to class "1" [47], [55].

It is important to note that the decision trees are generated in a very similar way to the algorithm described previously. Instead of node selection being made among all available nodes to be branched, Random Forest selects the node between a random set of nodes [48], [55].

Finally, **Naive Bayes** is a simple supervised learning algorithm based on the Bayes' theorem. This is characterized by adopting the so-called conditional independence of class, that is, considers that the presence or absence of an attribute is independent of the presence or absence of another attribute. Naive Bayes assumes that each molecular descriptor independently contributes to the probability of the drug causing the adverse effects under study. This becomes advantageous since the algorithm does not require a large amount of training data to estimate the parameters required for classification [49].

[47] Random Forests Algorithm, by Michael Walker, 2013. Available in http://www.datasciencecentral.com/profiles/blogs/random-forests-algorithm, accessed last time in 27-05-2017

[48] 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier. Available in http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, accessed last time in 27-05-2017

[49] How To Implement Naive Bayes From Scratch in Python, by Jason Brownlee, 2014. Available in http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/, accessed last time in 27-05-2017

Thus, we have given a role = label 'C' attribute of classes c1, c2, ..., cn is an attribute vector 'a' corresponding to all other attributes, the conditional probability of a ci class may be expressed as [50]:

$$P(C = ci|a) = \frac{P(C = ci) * P(a|C = ci)}{P(a)} \tag{2.2}$$

Naive Bayes kernel is a type of Naive Bayes. This one, unlike Naive Bayes, allows the use of continuous numerical attributes, being used in non-parametric estimation techniques. This operator is then based on kernel density estimation [55].

Thus, the conditional probability P(a|C=ci) will be calculated as the estimate of the kernel density for class ci, such that:

$$P(a|C = ci) = f_i(a) \tag{2.3}$$

with

$$f_i(a) = \frac{1}{Nh} \sum_{n=1}^{N} Ki \frac{(a - a_n)}{h} \tag{2.4}$$

Being that:

-a_n represents a training point;

-$Ki(a, a_n)$ represents the kernel function;

-h represents a smoothing parameter;

-N represents the total number of classes;

'h' is usually called bandwidth (bandwidth) and is a parameter chosen by the programmer [57].

A **regression** task predicts numbers and not labels (as in classification). Regression algorithms include statistical Linear and Logistic Regression, Artificial Neural Networks (ANN) and others [43, 45, 49].

In the task of **Association Rules Discovery**, association rules are created, based on how often the items occur together. That is, the intention is to identify frequent correlations among items, structuring regularities and affinities between them, in order to help the understanding of patterns [42, 45, 58].

A **clustering** task is quite different from the previous ones. Clustering aims the identification of groups with in the data. Contrary to the classification, it is not necessary to categorize the records in advance (unsupervised learning). Clustering is a descriptive task and can be used in image processing, data analysis, pattern recognition and many other domains. Some of the most used algorithms in clustering are K-Means and K-Medoids. The K-Means algorithm accepts an integer k and partitions the data in k clusters. Then, the objective is to calculate the similarity between the objects analyzed and the center of each cluster, with the object being inserted in the

[50] Naive Bayesian, by Dr. Saed Sayad. Available in http://www.saedsayad.com/naive_bayesian.htm, accessed last time in 27-05-2017

cluster with the greatest similarity, that is, with a smaller distance. Finally, whenever a new element is inserted, the center of the cluster is recalculated. In turn, the K-Medoids is a variation of the former. This algorithm is concerned with the concept of the most central object of the cluster. The most well-known implementations are PAM (Partitioning Around Medoids) and CLARA (Clustering LARge Applications) algorithms. An alternative method to clustering is the Density-Based Methods. Given a set of points in some space, it groups together points with many nearby neighbors, marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). This type of method has the capacity to obtain more satisfactory results, in situations where the distribution of data values is more dense. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is defined as being a fairly common algorithm of this type [51], [41, 49, 52].

Data Mining Tools

In order to allow professionals from other areas to benefit from DM and in order to facilitate its use, tools have been developed. Throughout this report, we will cover four popular tools that have been widely used [43].

The **Waikato Environment for Knowledge Analysis (WEKA)** tool was implemented in 1997 in New Zealand. It is an open source software written in Java, which allows it to run on different platforms. It implements a series of machine learning algorithms which gives it the ability to perform several DM tasks, such as: classification, regression, clustering, association rules discovery and visualization. Its great advantage derives from the fact that it allows comparing techniques, verifying which one presents a lower error rate [52], [53], [54], [43, 48].

RapidMiner, developed in 2001, is an open source tool written in Java and a no-coding required software. It uses both descriptive and predictive techniques, which allows to make cost-effective decisions. RapidMiner has algorithms to construct models for predictive analysis and also enables the graphical visualization of the results. Its use is advantageous as it ensures that complex analyzes and resulting forecasts can be integrated directly into the customer's infrastructure [52], [43, 48].

In turn, **KNIME** was considered the tool with the best performance and graphical interface. This one is also open source and is written in Java and was developed with the purpose of finding patterns. To do this, it applies algorithms to the data. It is used primarily in data preprocessing.

[51] A Tutorial on Clustering Algorithms. Available in https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html, accessed last time in 30-12-2016

[52] 12 Data Mining Tools and Techniques. Available in https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques, accessed last time in 30-12-2016

[53] "Mineração de dados com WEKA, Parte 1: Introdução e regressão." Available in https://www.ibm.com/developerworks/br/opensource/library/os-weka1/, accessed last time in 30-12-2016

[54] "Weka 3: Data Mining Software in Java." Available in http://www.cs.waikato.ac.nz/ml/weka/, accessed last time in 30-12-2016

Its graphical interface allows users to not only create data streams, perform the analysis steps, but also analyze the results [52], [55], [43, 48].

Finally, **R** is a scripting language, written in C and FORTRAN and is available on platforms such as Windows or Linux. Lately, interest in this tool has been increasing. This is verified since R is a fundamental tool in the development of decision support and data analysis systems such as the execution of complex tasks involving programming. It can be used in linear and nonlinear modeling, classification, clustering, and graphical and data analysis. As a disadvantage, it has an unfriendly interface [52], [56], [43, 48].

Metrics used to evaluate models

In order to test how a DM classifier model behaves on unseen data and to assess its quality, it is extremely important to use good evaluation practices. Different metrics have been proposed to evaluate classification models. It is important to note that different metrics aim to evaluate different characteristics of the classifier. Data classification problems are usually divided into binary, multiclass, and multi-labelled classification. In general, the metric measurement tool aims at evaluating the classifiers' performance and effectiveness. **Accuracy**, **Precision**, **Recall**, **AUC** and **F-Measure** represent some metrics for classification tasks [57], [58], [59].

Accuracy (acc) and error rate (err=1-accuracy) are known to belong to the group of metrics most commonly used to evaluate the generalization of classifiers, either in binary classification problems or in multiclass classification problems. These have the potential to differentiate each single solution and choose for the optimal designated solution, produced by a particular classification algorithm, from a large solutions space. For this, acc metric is based on the proportion of correct predictions about the total instances, that is, its function is to measure how many times the classifier makes the correct prediction [57], [3, 59, 60]. The accuracy is defined as below:

$$\frac{tp+tn}{tp+fp+tn+fn} \tag{2.5}$$

where, ***tp*** - true positive for ***C***; ***fp*** - false positive for ***C***; ***fn*** – false negative for ***C***; ***tn*** - true negative for ***C***.

On the other hand, the error rate reflects the proportion of errors committed in the whole set of instances and uses the percentage of incorrect forecasts to evaluate the solution produced. The error rate is defined as below:

[55] Examining the KNIME open source data analytics platform. Available in http://searchbusinessanalytics.techtarget.com/feature/Examining-the-KNIME-open-source-data-analytics-platform, accessed last time in 30-12-2016

[56] What is R? Available in https://www.r-project.org/about.html, accessed last time in 30-12-2016

[57] Turi Machine Learning Platform User Guide. Available in https://turi.com/learn/userguide/evaluation/classification.html, accessed last time in 04-01-2017

[58] Testing and Validation (Data Mining). Available in https://msdn.microsoft.com/en-us/library/ms174493.aspx, accessed last time in 04-01-2017

[59] Model Evaluation. Available in http://www.saedsayad.com/model_evaluation.htm, accessed last time in 04-01-2017

$$\frac{fp+fn}{tp+fp+tn+fn} \tag{2.6}$$

where, ***tp*** - true positive for ***C***; ***fp*** - false positive for ***C***; ***fn*** – false negative for ***C***; ***tn*** - true negative for ***C***.

From this it is verified that these metrics complement each other. The use of these metrics has many advantages, of which the fact that they are easily calculable and understandable by the users and of being applicable to problems called multiclass and multi-label. Regarding the error rate metric, the metric accuracy presents some disadvantages: its simplicity can lead to suboptimal solutions. Accuracy does not differentiate classes (positive or negative) [52], [3, 59, 60].

Metrics **F-Measure (FM)** and **Area Under the ROC Curve (AUC)** are also known to show good performance as a discriminator in binary classification problems. AUC is used primarily for imbalanced dataset. Despite its effectiveness, AUC presents a high computational costs, especially when it is used to discriminate solutions in multiclass classification problems, which is a limitation in its use. The performance of a classifier is portrayed by the value of the AUC metric. For a perfect classifier, AUC is equal to 1 [3, 59, 61].

The F-Measure is defined as below:

$$\frac{2*p*r}{p+r} \tag{2.7}$$

where, ***p*** - precision; ***r*** - recall.

The AUC value can be calculated as below:

$$\frac{Sp-n_p(n_n+1)/2}{n_p*n_n} \tag{2.8}$$

where, ***Sp*** is the sum of the all positive examples ranked, while n_p and n_n denote the number of positive and negative examples respectively.

Contrary to these, the **Precision** and **Recall** metrics do not have the ability to opt for the optimal solution, since they only have one evaluation task (positive or negative class) and they prove to be less reliable than the AUC metric. Generally, these metrics are used together. Precision allows you to identify, among the actual predicted items by the classifier, which are actually true [60], [3, 59, 60]. The precision is defined as below:

$$\frac{tp}{tp+fp} \tag{2.9}$$

where, ***tp*** - true positive for ***C***; ***fp*** - false positive for ***C***.

Recall can also be called sensitivity and designates the proportion of actual positive cases that are correctly identified [60], [3, 59, 60]. The recall is defined as below:

[60] Classification Accuracy is Not Enough: More Performance Measures You Can Use, by Jason Brownlee, 2014. Available in http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/, accessed last time in 04-01-2017

$$\frac{tp}{tp+tn} \tag{2.10}$$

where, ***tp*** - true positive for ***C***; ***tn*** - true negative for ***C***.

The Table 2.1 below presents a summary of the relevant characteristics of the most used metrics for classification problems.

Table 2.1: Threshold Metrics for Classification Evaluations [3].

METRICS:	**EVALUATION FOCUS:**
Accuracy (acc)	• In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	• Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Precision (p)	• Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	• Recall is used to measure the fraction of positive patterns that are correctly classified.
F-Measure (FM)	• This metric represents the harmonic mean between recall and precision values.

Evaluation methodologies

The evaluation of the model is a crucial step in the DM. This step allows the estimation of the performance of the models, so that the models with the best characteristics and that best represent the desired one are selected [50].

Cross Validation (CV) is one popular method of evaluation of DM models. It allows you to test each sample of the data set and avoids overlapping test sets. **K-fold Cross Validation** is the most common type of CV. This involves dividing the data set into k approximately equal size folds. The model construction is done k times and at each time one of the folds is used as test set while the others are used as training set. It is important to note that before performing cross validation, the subsets are stratified. In DM, normally the number of folds is 3, 5 and 10. As a main advantage, it presents the fact of displaying accurate performance estimation. However, it reveals some disadvantages: overlapped training data; underestimated performance variance [50], [61], [62], [62]. The Figure 2.4 shows an example of the 2-fold Cross-Validation technique.

[61] Model evaluation, model selection, and algorithm selection in machine learning, by Sebastian Raschka, 2016. Available in https://sebastianraschka.com/blog/2016/model-evaluation-selection-part3.html, accessed last time in 04-01-2017

[62] How to Evaluate Machine Learning Algorithms, by Jason Brownlee, 2013. Available in http://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/, accessed last time in 04-01-2017

Figure 2.4: Example 2-fold Cross-Validation.

As can be seen by the figure above, the data is divided into 2, so 2 models are built. For each model, half of the examples are used for training (represented by brown color) and the other half of the examples are used for testing (represented by orange color).

In small data sets, k is equal to the sample size, n, that is, equal to the number of instances in the data. In this case CV takes the name of **Leave-One-Out Cross Validation (LOOCV)**. This procedure ensures better data utilization and does not involve random sub-sampling. At the same time, its use is advantageous, insofar as LOOCV presents an unbiased performance estimate. On the other hand, it is quite expensive computationally and does not allow stratification. For this reason, most of the time, it is only used when the sample sizes are small [53], [63], [64], [62]. The Figure 2.5 shows an example of the Leave-One-Out Cross Validation technique.

Figure 2.5: Example Leave-One-Out Cross Validation.

As can be seen by the figure above, this method uses only one of its examples for testing (represented by orange color), and all other examples as training (represented by brown color) . In this way, as many models are created as there are examples (data).

Another were known model estimation technique is the **Hold-Out** method. This is characterized by dividing the data set into two main subsets. These are called: training set and test set. Training set is used for the construction of the predictive model, the training stage. Sometimes the training set is split into a "large" training set and a small validation set. Validation set is usually used to prune the model and/or to tune parameters of the algorithm. Finally, the test set allows

[63] Cross Validation, by Jeff Schneider, 1997. Available in https://www.cs.cmu.edu/ schneide/tut5/node42.html, accessed last time in 04-01-2017

[64] Evaluating what's been learned. Available in http://www.cs.ccsu.edu/ markov/ccsu_courses/DataMining-9.html, accessed last time in 04-01-2017

estimating not only the error rate of the classifier, but also the future performance of a model. The Hold-Out method is easy to compute the train and test set and it presents independent training and testing. However, it presents as a strong limitation that its evaluation can be highly varied, that is, the evaluation performed may differ, since it does not use all available data and depends on the way the training/testing division is made. Contrary to this, where some samples may never become part of the test set, k-fold CV ensures that all samples are used for validation, which makes this technique better than hold-out. It is concluded, therefore, that hold-out limitations can be overcome by k-fold CV [50],[54], [62].

The Figure 2.6 shows an example of the Hold-Out Cross Validation technique.

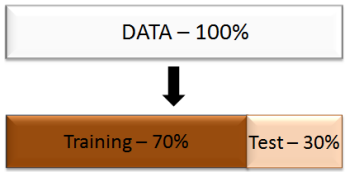

Figure 2.6: Example Hold-Out Cross Validation.

As can be seen by the figure above, the data set is divided into two subsets: training and test. Typically, the training set is larger than the test set. In this case, 70% is for the training data (represented by brown color) and 30% is for the test data (represented by orange color).

Another existing procedure that tries to overcome the limitations of Hold-Out, in very small data sets, is called **Bootstrapping**. Bootstrapping is a resampling method with replacement to form the training set and is used with great potential, especially for small samples. It has the ability to quantify the uncertainty associated with a given estimator. The best known bootstrap estimator is designated 0.632 bootstrap [55], [65], [62].

Limitations of Data Mining

As already mentioned, DM is characterized by its enormous potential in the process of knowledge extraction. In particular, in the area of health, the contribution of DM is indisputable, since its use may allow the identification of new disease patterns, making it possible for professionals to make rapid decisions. Despite all the benefits derived from its use, this method is not infallible and thus also presents some limitations that must be taken into account [43, 49, 50].

DM creates relationships and discovers patterns, yet does not necessarily explain them. Wang et al. [43, 49, 50] divides these limitations into four categories: statistical, data accuracy and standardization, technical and organizational.

It is essential to adapt the tasks and algorithms to be used for the intended purposes. In order for the results to be well interpreted, it is necessary that the relations established between the attributes be well defined. The number of variables used should be large. The success of using this

[65] Cross Validated. Available in http://stats.stackexchange.com/questions/103459/how-do-i-know-which-method-of-cross-validation-is-best, accessed last time in 04-01-2017

method depends greatly on the collected data, since the quality of the results obtained in the DM will depend directly on the quality of data collection and organization. At the same time, a high knowledge of the users is necessary [40, 43, 50].

Recommender Systems

The amount of digital information available on the Internet is enormous. This information overload is a major problem for users, since it makes it more difficult for them to access the items that interest them. In an attempt to minimize this setback, were created the Recommender Systems (RSs) in mid-1990s. This strategy has the ability to predict what users prefer based on their profile, their previous searches, and searches performed by users with similar interests. RSs have as their main objective to filter information and provide researchers with only relevant and highly correlated information. Thus, this method should provide adequate recommendations and suggestions, helping and directing users to the items that, from the outset, better serve their needs [1, 63, 64].

Currently, RSs play a key role on websites such as Amazon (https://www.amazon.com/), because these systems can also be useful and effective in increasing sales, since in suggesting products, they awaken the need and desire of the person to acquire them. These are also applied to recommendations for movies, music, travel services, among others. The development of a RS should take into account its purpose. Thus, in order to select the ideal algorithm and to promote an effective man-computer interaction, it is necessary to consider the specific requirements, specifications and limitations. Most of these use various algorithms of machine learning and data mining [1, 65].

RSs are a multidisciplinary field and can take two main groups:

-Item Recommendation: can further be divided into content-based recommendation.

-Item Rating Prediction: can further be divided into recommendation with collaborative filtering.

The must relevant approaches will be presented. Is necessary understanding the features of different recommendation techniques. Figure 2.7 shows the anatomy of different recommendation filtering techniques.

Collaborative filtering (CF) is an approach that designates a recommendation algorithm known to base its predictions on previous behaviors and opinions of other users with similar characteristics and interests, in order to simulate user preference. Thus, the recommendations made to the active user are based on items that other users with similar choices performed in the past. For such, the CF builds a database of user preferences, then compares users with identical tastes, then calculates the similarities between them and generates the recommendations. This technique is considered the most popular and most implemented and subdivided into two categories: memory-based and model-based. The first is obtained through user-based techniques and items. The second is characterized by the construction of a model through previous classifications and its purpose is

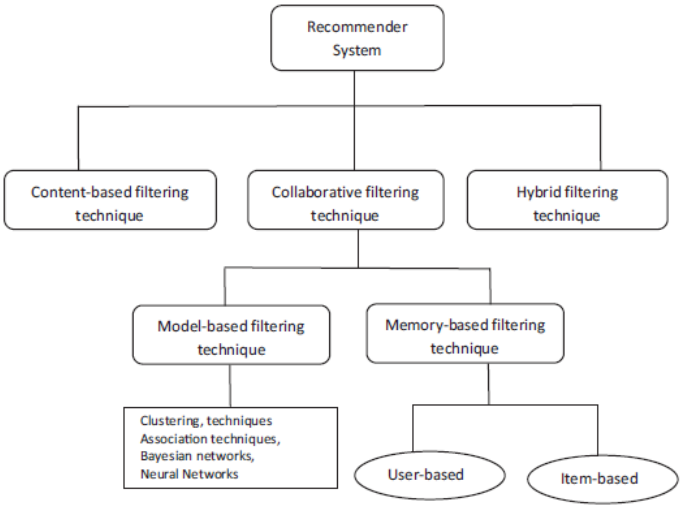

Figure 2.7: Recommendation techniques [1].

to solve the dispersion problems associated with RSs. For this, it uses association techniques, Clustering, Bayesian networks and decision trees. Association rules mining algorithms present the compressed data in order to improve both storage efficiency and performance. Clustering is used, for example, in image processing, statistical data analysis or pattern recognition, and aims to divide a set of data into a set of sub-clusters, thus reducing the set of candidates in collaboration-based algorithms. Consequently, the recommendations are made based on the average of the participation clusters. In the clustering method, K-means and Self-Organizing Map (SOM) are known to be the most used. On the other hand, when user preferences change slowly over time and, therefore, time is taken to construct the model, Bayesian classifiers are used. Finally, to classify examples that had not yet been seen and to handle items with missing characteristics, we used the construction of a tree graph, called a decision tree. This method is more understandable than techniques such as SVM and Neural Networks. Despite the advantages implicit in the use of CF, such as: ability to recommend items in areas where there is not much content and produce, in most cases, acceptable results; CF presents some limitations such as the fact that it requires not only a large number of user feedbacks but also that users have purchased the same product in order to create standards for recommendations [66], [1, 63, 65].

The algorithms exploited in this dissertation were those of predictive rating with collaborative filtering. The extension used in this dissertation - Recommender extension - provides 10 options of training algorithms for forecasting, being that, in this project, 3 were addressed: **Matrix Factorization**, **Slope One** and **User k-NN**.

The **matrix factorization** has proved to be a promising methodology for collaborative filtering recommenders, since it allows modeling various real-life situations. Recommendation systems depend on different types of input data. These are generally presented in the form of a matrix

[66] Introduction to approaches and algorithms, 2013. Available in https://www.ibm.com/developerworks/library/os-recommender1/, accessed last time in 22-12-2016

with a dimension representing the users and another representing the items. It allows essentially characterizing both items and users by factor (k) vectors inferred from item classification patterns. This approach leads to a recommendation when there is high correspondence between item and user factors [67].

The matrix R is considered to result from a set (in this project) of Drugs - D and a set of ADR's - A, where D is the users and A are the items. Thus, the matrix R of size |D|*|A| has all the rating's representative of the Drugs-ADR ratios. This recommendation algorithm intends to find the matrices P (|D|*k) and Q (|A|*k) such that its product approaches the matrix R:

$$R \approx P * Q^T \tag{2.11}$$

The forecast of a rating $(m_i - a_j)$ is obtained by calculating the scalar product of the two vectors corresponding to the drug d_i and the adverse effect a_j:

$$r_{u,i} = d_i^T a_j = \sum_{k=1}^{k} d_{ik} a_{jk} \tag{2.12}$$

To obtain the P and Q matrices, the system minimizes the regularized quadratic error for each Drugs-ADR pair between the predicted values and the set of known ratings:

$$e_{ij}^2 = (r_{ij} - \sum_{k=1}^{k} p_{ik} q_{jk})^2 \frac{\beta}{2} \sum_{k=1}^{k} (||P||^2 + ||Q||^2) \tag{2.13}$$

Note that the B parameter used in the formula is to prevent the model from overfitting. To do so, this parameter attempts to control the magnitudes of the vectors so that P and Q are a good approximation of R without containing large numbers [68].

Proposed in 2005 by Daniel Lemire and Anna Maclachlan, **Slope-One** represents an item-based collaborative filtering algorithm for easy and efficient implementation. This algorithm aims to solve the problem of data dispersion, generating recommendations, that is, it is applied when it is necessary to fill the vacancies of the item-user matrix. Based on linear regression, it calculates the mean difference between the scores of the two items [66, 67, 68].

Given a training set X, and any two items j and i with ratings uj and ui, respectively, the mean deviation between item i and item j is taken as:

$$dev(j,i) = \sum_{u \in S_{j,i}(X)} \frac{u_j - u_i}{card(S_{j,i}(X))} \tag{2.14}$$

To do that, this algorithm only considers the users who have classified the target items and also considers the other items that have been evaluated by those same users. Note that card represents the cardinality (number of elements in the set) [69, 70].

[67] MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS, by Yehuda Koren, Robert Bell and Chris Volinsky, 2009. Available in http://www.usabart.nl/PTQR/sameer1.pdf, accessed last time in 27-05-2017

[68] Matrix Factorization Techniques For Recommender Systems, by Markus Freitag, Jan-Felix Schwarz, 2011. Available in https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/lehre/SS2011/Collaborative_Filtering/pres1-matrixfactorization.pdf, accessed last time in 27-05-2017

After calculating the average deviations, the forecast is calculated:

$$P(u)_j = \frac{1}{card(R_j)} \sum_{i \in R_j} (dev_{j,i} + u_i) \tag{2.15}$$

where, $(R_j = i | i \in S(u), i \neq j, card(S_{j,i}(X)) > 0)$ is the set of all relevant items.

User k-NN (k-Nearest Neighbor) designates a collaborative type of user-based filtering that determines the grouping of these based on their similarities. Thus, it is verified that this algorithm foresees a given rating of user "a" to item "i", calculating the degree of similarity that exists between that user "a" and other users, that is, the "nearest neighbors" of user "a". Such a degree can be obtained through the similarity of Cosine:

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.16}$$

or Pearson's Correlation:

$$\frac{\sum_{i=1}^{n} (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n} (A_i - \bar{A})^2} \sqrt{\sum_{i=1}^{n} (B_i - \bar{B})^2}} \tag{2.17}$$

Being that:

-$(A_i B_i)$ correspond to the ratings of user's item 'a' and 'b', respectively;

-n is the total number of items sorted by users 'a' and 'b';

-$(\bar{A}\bar{B})$corresponds to the average of all user ratings 'a' and 'b', respectively.

When user "a" is compared with other users, only the k-NN, that is, the closest/closest neighbors are saved. This avoids overfitting the model. After this comparison, the user's rating "a" to item "i" results from multiplying the average of the ratings of the neighboring k's most similar to the user 'a' by the measure of similarity between the user 'a' and the neighbor in question [69], [71].

Another widely used approach today is called **Content-Based Filtering (CBF)**. This is characterized by, based on the profile, on the characteristics and behavior of the user, build a recommendation for that user. That is, this technique depends on the metadata of the items and enhances the recommendation of items similar to those that the user liked previously. Unlike CF, CBF ignores the choices of other users. Like the above mentioned technique, there are some disadvantages which arise in requiring a descriptive and organized knowledge of the user's choices and preferences before a recommendation is made for this. At the same time, its effectiveness is compromised, as it depends on the availability of user data. Finally, CBF faces the fact that it only suggests items similar to those already chosen by the user, which can be a problem, since the user may no longer be interested in this type of products. However, on the other hand, this type of filtering can, within a short time, adjust its recommendations to new user preferences, which is one of its main advantages [58], [1].

In order to overcome the limitations of the previously mentioned techniques and to increase the effectiveness and precision of RSs, the **hybrid approach** emerged, resulting from the combination

[69] K-NN (RapidMiner Studio Core), by RapidMiner, 2017. Available in https://docs.rapidminer.com/studio/operators/modeling/predictive/lazy/k_nn.html, accessed last time in 27-05-2017

of CF and CBF. This technique tries to use the advantages of CF to appease the disadvantages of CBF and the opposite also happens. It follows that the use of combined algorithms, in most cases, is more favorable and has better results than the use of individual algorithms [58], [1, 63, 65].

2.3 Ontologies and standards

It is common practice by both research organizations and companies to store large amounts of information. That enormous amount of information and data has to be selected, organized and treated. For each domain there are terms, definitions, relations, etc that can be used and interpreted differently by the community. Thus, there is the need to develop ontologies in order to standardize concepts, create norms and cross data, allowing the establishment of coherence and consensus among this community. These ontologies play a crucial role, since they can be considered as a formal representation of knowledge, whose main objective is to facilitate communication and sharing between community and to improve an existing information organization [70].

In 1993, Tom Gruber [72] defined ontology as "a specification of a conceptualization, or a description of concepts and relationships that exist between these concepts". Years later, in 1997, Borst [72], on the other hand, classified ontology as "a formal specification of a shared conceptualization". Thus, in general, it is considered that the ontologies, created by specialists, specify concepts, define terms and relations that are essential to model a domain. The concept of ontology is ambiguous and depends on the context in which it is being used. Usually these are a set of classes organized in a hierarchy. In the context of computing, more specifically, in database systems, the ontological term is used for integrated heterogeneous data banks, enabling interoperability between different systems and specifying interfaces for independent knowledge-based services [70], [73, 74, 75].

The use of ontologies promotes a clearer organization of information. At the same time, they allow for better communication and information sharing among people, so as to promote coercion, clarity and understanding about a domain, eliminating inconsistencies in the representation of knowledge. On the other hand, despite its remarkable relevance, there are some drawbacks that can be considered. It is necessary that the ontologies be updated and evolve and not go through a process of stagnation. For this, it is fundamental to have found tools for the construction of the same and that are described methodologies for its development. Finally, there is a large number of languages to encode, which makes it impossible to exchange data between systems and a transfer of knowledge they do not share the same languages [73, 74, 75].

In this section, three types of ontologies are addressed: Ontologies for Medical application, Ontologies for Drug Reactions and Ontologies for Chemoinformatics.

[70] Ontologies and Semantic Web. Available in http://www.obitko.com/tutorials/ontologies-semantic-web/what-is-ontology.html, accessed last time in 13-12-2016

2.3.1 Encoding Ontologies

Ontologies are formalized using ontological languages, that is, formal languages. Ontology languages enable the encoding of terms from specific domains and rules of reasoning that support the processing of knowledge. Ontological languages are declarative by nature. Most of the time, they are generalizations of framework languages and are based on first-order logic or on description logic [71], [76].

The Web Ontology Language (OWL) is an indispensable technology for the future implementation of the Semantic Web. OWL has been defined as a language that allows not only defining, but also instantiating, ontologies on the Web. It enables the processing of information content and interpretation by machines. Because it is based on eXtensible Markup Language (XML), it facilitates the exchange of information between different types of computers, using programming languages [71], [76].

There are several traditional languages of syntax ontology, among which we can highlight: Common Logic, F-Logic, Developing Ontology-Grounded Methods and Applications (DOGMA), KM programming language, among others [71].

2.3.2 Domain Ontologies

Ontologies for Medical application

In order to facilitate and standardize communication between a medical community and improve understanding about human diseases, the Disease Ontology (DO) has been developed. This intends semantically integrate medical vocabularies, thus allowing coherence in the description of terms related to human diseases. These terms are related to well-established terminologies, such as SNOMED [72], which contain not only human diseases, but also concepts related to them. This open source ontology allows the crossing of data between symptoms, signs, and genetics or infectious diseases [73].

Ontologies for Drug Reactions

Thus, in the attempt to standardize and integrate data in the domain of adverse reactions caused by medical interventions, whose objective is to alleviate pains or prevent diseases, the Ontology of Adverse Effects (OAE) has been created. This biomedical ontology, based on the community, is also aimed at computer-aided reasoning [74], [75], [2, 77]. In OAE, an "adverse event is defined as a pathological bodily process that is induced by medical intervention" [72], [2, 77].

[71] Web Ontology Language, 2016. Available in https://en.wikipedia.org/wiki/Web_Ontology_Language, accessed last time in 15-01-2017

[72] SNOMED International SNOMED CT Browser. Available in http://browser.ihtsdotools.org/?, accessed last time in 15-01-2017

[73] Disease Ontology, 2016. Available in http://disease-ontology.org/, accessed last time in 09-12-2016

[74] Ontology of Adverse Events. Available in https://bioportal.bioontology.org/ontologies/OAE, accessed last time in 10-12-2016

[75] AE. Available in http://www.oae-ontology.org/, accessed last time in 10-12-2016

To better understand the mechanisms of induced adverse reactions and to facilitate the organization of the information and the existing content, one appeals the aid of diverse resources of vocabulary, such as the MedDRA, the World Health Organization-Adverse Reaction Terminology (WHO-ART) and the Common Terminology Criteria for Adverse Events (CTCAE). MedDRA is known a vocabulary of codification of adverse events preferred by the FDA and used in such a way for the Vaccine Adverse Events Reporting System (VAERS), as for FDA Adverse Event Reporting System (FAERS). In turn, the WHO keeps dictionary WHO-ART with the goal that this serves of base for the rational codification of terms of adverse reaction. Finally, the USA National Cancer Institute (NCI) developed vocabulary CTCAE so that he is used in the evaluation of ADRs associated with cancer therapy drugs. These vocabularies intend, from individual data, symptoms or illnesses, to distinguish the common characteristics and aspects not perceivable [72], [73], [2, 77].

The OAE presents a hierarchic relation of terms. There is a database called ADReCS which provides the standardization and hierarchical classification of four levels of ADR terms. As shown in Figure 2.8, the adverse effect of different types are divided in subclasses and classified in accordance with the type of medical intervention [2, 77].

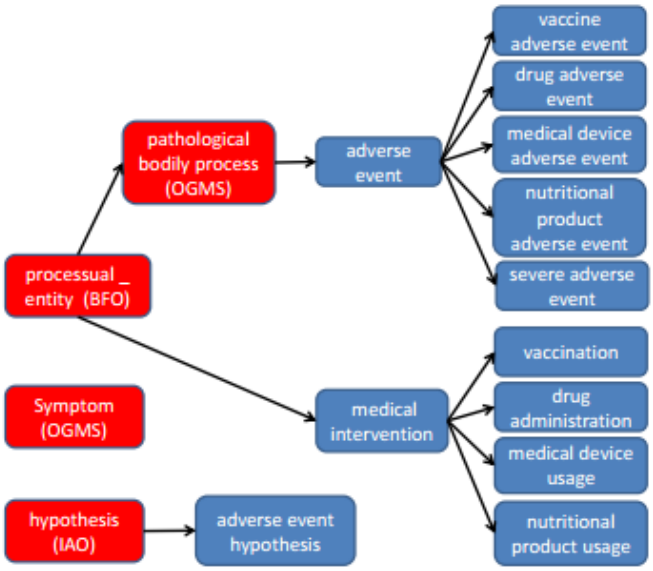

Figure 2.8: Key terms in OAE [2].

Currently, this ontologie presents 484 representational units, written down by means of 369 terms with specific AEO identifiers and 115 terms with specific OAE identifiers [2, 77].

Ontologies for Chemoinformatics

As is known, the amount information available in chemistry domain has been increasing rapidly. This presupposes a more efficient and coherent organization of data. In order to filter, retrieve and elucidate information, it is essential, through computational processing, to hierarchically organize

knowledge. Thus, the ontologies have been used in order to promote a better management of data in large volume and to allow a hierarchical classification concise [78, 79, 80].

The Chemical Entities of Biological Interest (ChEBI) ontology constitutes a database, which is freely available (https://www.ebi.ac.uk/chebi/), and can be used for purposes such as text mining. It is responsible for grouping chemical entities, such as atoms, molecules, and creating similarity patterns. Based on structural chemical characteristics, biological functions and system interactions, ChEBI provides a nomenclature and a hierarchical classification of the molecules [78, 79, 80].

The hierarchical classification of molecules has many advantages. Allowing not only narrowing in the area of interest within a large domain, but allowing a compact representation of the generalized knowledge at the highest level to which it applies, therefore, a search and access to people become easier. In addition, the hierarchical organization also plays a key role with respect to useful predictions and the prediction of properties of new entities [78, 79, 80].

It is important to emphasize the importance of the ontologies for the molecules, since an interpretation of biological mechanisms can allow the detection of bioactivity patterns associated with the chemical structure of the molecules [78, 79, 80].

2.4 Related work

After conducting an in-depth research on the subjects and subjects in question and being fully aware of them, it was extremely necessary and useful to carry out a search for the work that has already been done in this area. The main purpose of this research is based on the fact that it is important and necessary to know what has already been done, what methods/tools/strategies have been used, what conclusions have been obtained, the progress achieved, the problems faced, what can be improved and, essentially, which has not yet been done and which could bring great progress in this area. This will ensure potential and effectiveness of the technology for predicting ADRs during drug development.

It has been verified that some researches have already been done that use Data Mining algorithms and tools for this purpose. As an example, a study by Diogo Pinto et al. [81]. In this investigation, the main objective was focused on the use of algorithms and, in their subsequent evaluation, for recommendation systems. For this, data from the ADReCS database were used. The results were satisfactory in that Diogo Pinto et al. was able to show that the technology used allowed the forecast of ADRs, mainly in the pre-commercialization phase.

All the investigations carried out, even those that did not observe such satisfactory results, are extremely useful and relevant to future conclusions. It is important to test, use different methodologies, experiment with tools, make observations, continue studies and conduct research, so that techniques can be applied successfully and above all, can bring benefits and safety to the health of the world population.

2.5 Chapter Summary

This chapter it was presented the relevant basic concepts on the domain of adverse drug reactions and Data Mining.

The relevant Web repositories to be used in the study were also identified. The survey on the Data Mining algorithms have also helped to make a decision on which ones to use. As is usual on scientific studies, methodologies and metrics were investigated and established to evaluate the experiments this project.

It is important to remember that ontologies are important and useful in that they intend to standardize concepts and create norms, in order to establish a consensus on the available information. This allows a better organization of the same, which leads to a clearer and more precise communication/sharing of knowledge. There are several types of ontologies, and in this work only those that may have interest in the future work were addressed.

The ontology survey made in this chapter was useful to assess whether the use of additional information in the field of chemistry can improve the performance of data mining algorithms.

Chapter 3

Experimental evaluation

The dissertation work is composed by a set of Data Mining experiments to answer the two research questions stated in section 1.2. The case studies of this project used data sets from the ADReCS and DrugBank databases. The data and information available in these databases were pre-processed to make them adequate for the analysis tools. Subsequently, with the aid of enrichment tools, namely PaDEL and Open Babel, the processed data were enriched, in order to perform ADR prediction on new drugs. Lastly, the data were transformed into a suitable format for the DM methods and tools and finally, models were constructed (see Figure 3.1).

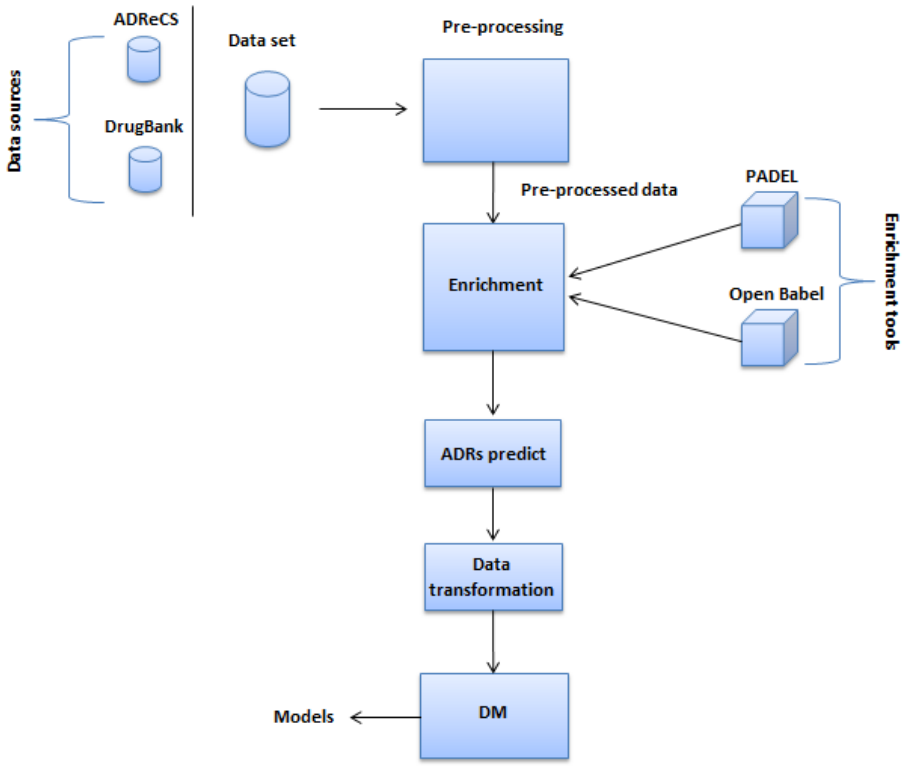

Figure 3.1: Web repositories, workflow of the analysis tasks and tools involved in the research work.

In the experimental part of our work, we started by replicating Diogo's work [81] previously discussed in 2.4. We have used recommendation systems as Diogo and then, evolved to classification tasks. In this set of experiments, the ADReCS database (described in 2.1) was used. The table 3.1 shows the methodology for this. Recommendation system was used as predictive methods only, whereas the classification can provide potential explanations for the ADR apart from the predictive facility.

Table 3.1: ADR prediction with Recommender Systems using ADReCS.

Methodology: ADR prediction with Recommender Systems using ADReCS

ADRsDataSets = {WholeADReCSARDs & 1LevelUpInHierarchyOfADRs }
  /* 1LevelUpInHierarchyOfADRs is the set of ADRs
  in the hierarchical level just above the leaves */
RSalgorithms = {LDA }

Forall DS in **WholeADReCSADRs 1LevelUpInHirarchyOfADRs**
  Forall Alg in **RSalgorithms**
    build and evaluate models
  EndFor
EndFor

For the classification tasks, a previous data enrichment stage was performed. The enrichment was done with addition of molecular descriptors. It was evaluate if the amount of drug information improves both the predict made but also the potential explanations given (implicit in the models). That is, it was observed to what extent the use of more information can be more useful and advantageous. The table 3.2 below shows the methodology for this.

Table 3.2: ADR prediction experiments with classification algorithms using ADReCS.

Methodology: ADR prediction with Classification Algorithms using ADReCS repository

DMalgorithms = {RandomForest & SVM & k-NN & NaiveBayes }
Forall DS in **WholeADReCSADRs 1LevelUpInHirarchyOfADRs**
  find molecules of the active principle of the drugs
  Enrich DS with Molecular Descriptors
  Forall Alg in **DMalgorithms**
    build and evaluate models
  EndFor
EndFor

The Data Mining tool chosen to carry out the process was RapidMiner, since it has an easy to use interface and enables the definition of the Data Mining's steps workflow. The version of RapidMiner used was educational one, since it allows free and unrestricted access of the number of lines to students and teachers.

After the analysis of ADReCS data, we have investigated the problem of drug to drug interaction. For that purpose, we have used the DrugBank database. DrugBank (described in 2.1) is

a freely available database. It is considered a unique biology and chemistry resource that contains detailed data not only of drugs, but also of the amino acid sequence, structure and route of administration [1].

Briefly, this project focused on two case studies. In the first case study, two experiments were made with the ADReCs database data. The first experience focused on the prediction of existing drug-adverse effects pairs. Thus, we only used information from the databases and used algorithms of recommendation systems that was responsible for the prediction. The recommendation algorithms used were Matrix Factorization (MF), Slope-One and User k-NN. Models performance were assessed using obtaining the metrics: Accuracy, Precision, Recall and F-Measure. The second experiment was mainly intended to find not only drug-adverse effects relationships, but also biochemical justifications for their existence. To this end, classification methods were used as described in 2.2. The classification algorithms used were Decision Tree with CART, Random Forest with CART, Naive Bayes and Support Vector Machine (LibSVM). In addition to the available information in the databases, we have added data related to the molecular descriptors of each drug. The performance of the model was measured by obtaining the metrics: Accuracy, Precision, Recall and F-Measure. The last case study involved only the pre-processing of the data taken from DrugBank.

Therefore, this chapter essentially addresses the relevant information regarding the content of the data used and describes in detail all the steps involved in the realization of the experiments.

3.1 Case Study 1

3.1.1 Experiment 1

Data pre-processing

This case study used data sets from the ADReCS [2] (described in 2.1). This database is only available in XML format, and was converted to the CSV format, format accepted by the RapidMiner tool.

As is usual, the original data was preprocessed in order to select only the most relevant one for the study. The data reduction process performed in the data preprocessing phase is very useful. This allows to reduce the training time of each algorithm and improve the model quality, since infrequent attributes could lead the model to make noise-based decisions, thus causing overfitting of the system.

In this experiment, we have used the leaf nodes of the ADR hierarchy. The original data contained about 1353 drugs and 6722 ADR's. In order to reduce not only the drugs but also the less frequent adverse effects, the data was filtered. Thus, all drugs with less than 100 associated adverse effects were eliminated. In parallel, all adverse effects that occurred less than 3 drugs were also eliminated.

[1] DrugBank. Available in https://www.drugbank.ca/about, accessed last time in 09-06-2017
[2] http://bioinf.xmu.edu.cn/ADReCS/

At the end of the data pre-processing, we obtained a file with 615 drugs and 3058 adverse effects. The CSV format of the data can be seen as a matrix where each line has information concerning a single drug and each column has information concerning a single ADR. Apart from first line and first columns that have names, each cell I, j stores a 0 if there are no ADR caused by line I drug and a 1 if drug in line I causes ADR at column.

Recommendation Process

In order for the RapidMiner recommendation process operators to apply the recommendation systems discussed in 2.2, it was necessary to use a name recommender extension (http://www.e-lico.eu/ Recommender-extension.html) provided by RapidMiner. The algorithms exploited were those of predictive rating with collaborative filtering.

The general recommendation process achieved is presented in the Figure 3.2.

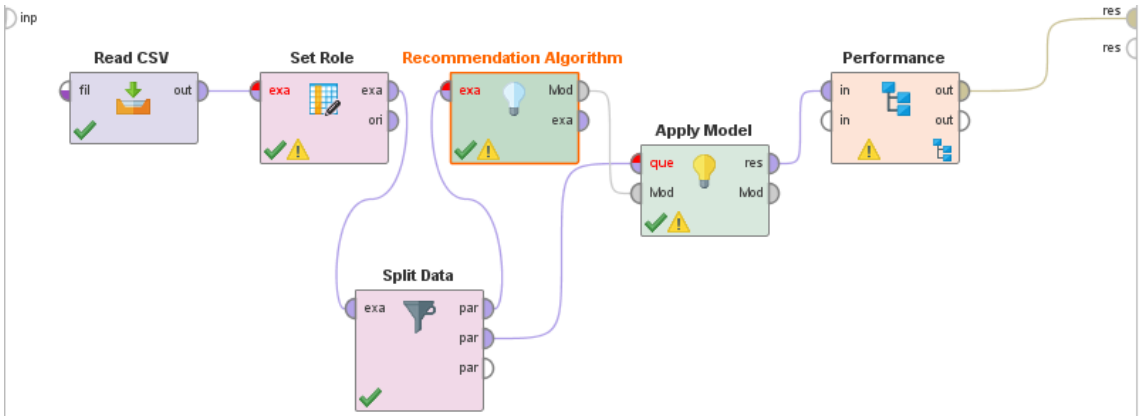

Figure 3.2: Recommendation Process.

Note that the "Read CSV" operator was first used in order to read the CSV file obtained in the pre-processing phase described previously in 3.1.1. Figure 3.3 shows a sample of data with 3 attributes: Drug, ADR, Rating.

ExampleSet (1880670 examples, 0 special attributes, 3 regular attributes)

Row No.	Drug	ADR	Rating
1	BADD_D00001	15.05.03.016	1
2	BADD_D00001	22.04.03.003	0
3	BADD_D00001	17.19.01.001	0
4	BADD_D00001	16.29.01.001	0
5	BADD_D00001	12.02.05.007	0

Figure 3.3: Sample of data.

Later, in order to relate each of the three existing attributes - Drug, ADR, Rating - with Role predefined by the extension used - "user identification", "item identification" and "label" - the operator "Set Role" was used. As can be seen in the following figure (Figure 3.4), the drugs were considered the users and the adverse effects were considered items/content.

attribute name	target role
Drug	user identification
ADR	item identification
Rating	label

Figure 3.4: Parameters used in the "Set Role" operator.

"Split Data" operator was next used. This operator partitions the input data set into subsets according to the specified relative sizes. The "Recommendation Algorithm" operator is the next in the analysis workflow. In order to use recommender systems, we had to install an extension to RapidMiner. The RapidMiner extension used provides 10 options of training algorithms for forecasting. In this project, we have used 3: Matrix Factorization, Slope One and User k-NN (described in 2.2). Next in the workflow is the "Apply Model" operator. It is responsible for receiving the training model, that is, it receives the model constructed using the training set. The test file, which contains the Drug-ADR pairs whose rating will be predicted is also provided. In the output, we have a file with one more attribute, namely "prediction" attribute. The last sub process, called "Performance", compotes the metrics Accuracy, Recall, Precision and F-Measure. These are presented as a percentage. The higher the percentage, the better the model performance. This sub process comprises 6 operators (Figure 3.5). The first one, called the "Set Role", was applied in a way to convert the Role of the regular prediction attribute to prediction.

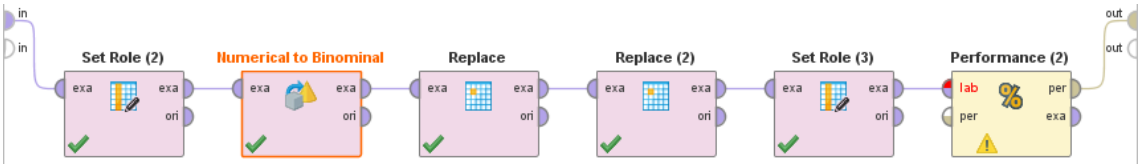

Figure 3.5: "Performance" sub process.

Then, the operator "Numerical to Binomial" was used to transform the prediction attribute to the binomial type. These two conditions were necessary in order to be able to subsequently apply the operator "Performance Binomial Classification". This resulted, on the part of the prediction attribute, from a transformation of examples <0.5 in false and examples >= 0.5 in true. In order to compare the "prediction" attribute with the "rating" attribute, the examples of the prediction attribute were converted from true to 1 and from false to 0. This was accomplished with the two operator "Replace". The "Set Role" operator (Figure 3.6) was applied again in a way to convert the Role of the prediction attribute to prediction.

attribute name	target role
Drug	user identification
ADR	item identification
Rating	label
prediction	prediction

Figure 3.6: Parameters used in the "Set Role (3)" operator.

In the Figure 3.7, it can be seen a sample of the output file.

Row No.	Rating	prediction	Drug	ADR
1	1	1	BADD_D00001	15.05.03.016
2	0	1	BADD_D00001	22.04.03.003
3	0	1	BADD_D00001	17.19.01.001
4	0	1	BADD_D00001	16.29.01.001
5	0	1	BADD_D00001	06.01.01.004

Figure 3.7: Sample of the output file.

Finally, the operator "Performance Binomial Classification" was applied. This allowed obtain the accuracy, recall, precision and F-Measure metrics (detailed in 2.2).

To assess the robustness of the methods, we have produced two data sets by introducing two different noise-levels in the original data. The data sets used was prepared with 5% and 10% noise and the procedure previously described was performed 2 times for each noise level. The effect of noise on the value of the accuracy metric and, consequently, in the model performance was recorded.

3.1.2 Experiment 2

Data pre-processing

In this experience, we have used 26 groups of ADRs ids. ADRs ids are organized in an hierarchy and we have merged the levels of the hierarchy into the top level ids of the hierarchy. That is, the previous ADR has become an ADR of group 3, identified by the first 2 digits of the identifier. The Figure 3.8 present the 26 nodes, corresponding to the 26 groups of adverse effects explored in the present experiment.

Taking into account that the results matrix, contrary to the results matrix of previous experiment, is quite dense, it was necessary to eliminate the drugs whose adverse effects cover many of the groups of adverse effects presented previously.

The database was enriched with a set of attributes called molecular descriptors, which qualifies each drug relative to its typology, molecular constitution, among other properties. To do this, all

01 Blood and lymphatic system disorders (1091 drugs)
02 Cardiac disorders (1405 drugs)
03 Congenital, familial and genetic disorders (280 drugs)
04 Ear and labyrinth disorders (715 drugs)
05 Endocrine disorders (763 drugs)
06 Eye disorders (1030 drugs)
07 Gastrointestinal disorders (1538 drugs)
08 General disorders and administration site conditions (1586 drugs)
09 Hepatobiliary disorders (829 drugs)
10 Immune system disorders (1368 drugs)
11 Infections and infestations (1200 drugs)
12 Injury, poisoning and procedural complications (1151 drugs)
13 Investigations (1301 drugs)
14 Metabolism and nutrition disorders (1301 drugs)
15 Musculoskeletal and connective tissue disorders (1265 drugs)
16 Neoplasms benign, malignant and unspecified (incl cysts and polyps) (489 drugs)
17 Nervous system disorders (1555 drugs)
18 Pregnancy, puerperium and perinatal conditions (182 drugs)
19 Psychiatric disorders (1251 drugs)
20 Renal and urinary disorders (1101 drugs)
21 Reproductive system and breast disorders (813 drugs)
22 Respiratory, thoracic and mediastinal disorders (1309 drugs)
23 Skin and subcutaneous tissue disorders (1566 drugs)
24 Vascular disorders (1528 drugs)
25 Surgical and medical procedures (329 drugs)
26 Social circumstances (241 drugs)

Figure 3.8: Groups of adverse effects at level 1 of the hierarchy.

the SMILES formulas to represent molecules was used (example: Aspirin - CC (= O) Oc1ccccc1C (= O) O), through a Java program that accesses the ChemSpider webserver [3].

Subsequently, we used the web server PaDEL Descriptors to obtain the molecular descriptors. It is important to note that in order not to reduce the number of drugs to use, we chose to ignore the 3D descriptors, since this platform only provides 3D descriptors for only 333 drugs in the ADReCS database.

Some attributes representing molecular descriptors have been removed because they produce a large number of missing values.

Next, we created the attribute that the algorithms will try to predict, that is, the "label" attribute. This was taken from the ADReCS database, using the adverse effects hierarchy. The "groupADR" attribute of role = label has taken values of true or false. When " groupADR" = true, it is assumed that the investigational medicinal product has adverse effects relating to the group of adverse effects in question. When “groupADR" = false, it is stated that the medicinal product in question is not responsible for causing any of the adverse effects relating to the group of adverse effects in question. This can be observed in a small sample below (Figure 3.9), which is related to group 1 of adverse effects - blood and lymphatic system disorders.

[3] Available in http: /www.chemspider.com/, accessed last time in 27-05-2017

Name	grupoADR	nAcid	ALogP	ALogp2	AMR	...	apol	naAromAt	nAromBon	nAtom
BADD_D00005	true	1.0	-1.102199	1.2148448	36.9694		23.342722	0.0	0.0	22.0
BADD_D00009	true	0.0	-0.231400	0.0535459	44.754900		22.785136	6.0	6.0	20.0
BADD_D00013	true	0.0	-1.194999	1.4280249	38.2306		25.692687	0.0	0.0	26.0
BADD_D00014	true	1.0	-0.499199	0.2492006	36.811800		21.207136	0.0	0.0	19.0
BADD_D00015	false	1.0	0.4399999	0.1935999	47.9988		24.382343	6.0	6.0	21.0
BADD_D00026	true	0.0	8.9349999	79.834224	189.2906		107.74040	0.0	0.0	96.0
BADD_D00029	false	0.0	-0.734799	0.5399310	106.25440		64.246996	0.0	0.0	57.0
BADD_D00038	true	0.0	-2.0427	4.1726232	33.4662		16.669171	9.0	10.0	14.0

Figure 3.9: Sample of the file corresponding to the groupADR=1.

After the data treatment, we obtained a data set consisting of "Name", "GroupADR" and 1524 molecular descriptors. This set contains 340 drugs. During the execution of the process, only 10 molecular descriptors were selected.

Classification Process

It is important to recall that this experiment is mainly intended to find not only drug-adverse effects relationships, but also biochemical justifications for their existence. For that, four classification algorithms were used: Decision Tree with CART, Support Vector Machine (LibSVM), Random Forest with CART and, finally, Naive Bayes (detailed in 2.2). Taking into account the nature of the molecular descriptors used in this set of experiences, it was necessary to resort to the "Naive Bayes (kernel)" operator. Each of these classification algorithms was executed 26 times, which corresponds to the number of adverse effects groups in the study. Each trial attempts to predict whether or not the medicinal products have adverse effects from the group of adverse effects under study.

The flow chart representative of the classification process is presented in the Figure 3.10.

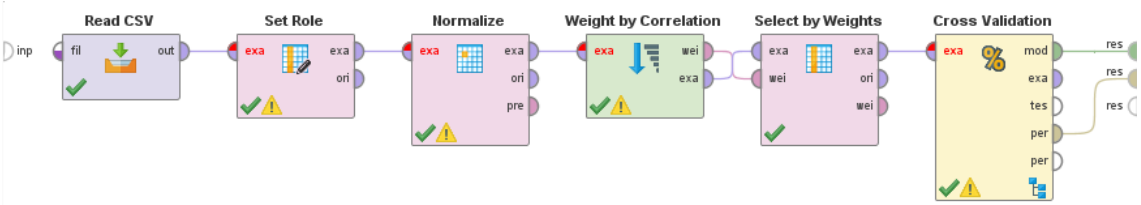

Figure 3.10: Classification Process.

As can be seen, this process was started with the "Read CSV" operator. This operator is responsible for reading the file. The input file used was that obtained in data pre processing (described in 3.1.2). This includes a "Name" attribute, a "groupADR" attribute, and a set of attributes representative of the molecular descriptors of each drug.

After reading the file, it was necessary to change the role of the attribute "groupADR" from "regular" to "label", so that the use of any classification algorithm was possible. To do this, the

"Set Role" operator was applied. It is important to emphasize that it is the classification algorithm that makes the forecast according to this attribute.

We have then applied the operator "Normalize" to all attributes in the database. This was fundamental to transform all the attributes, in order for all to have the same range of values. That is, it was intended to prevent attributes with large values would have a greater influence in the forecast.

Next, the most important attributes for the study were identified. Least important attributes were discarded. This feature selection was achieved with the help of the "Weight by Correlation" and "Select by Weight" operators. The first one, calculates the importance (weight) of a given attribute. This weight corresponds to the correlation between each input attribute and the label. We should mention that the greater the weight of a given attribute, the greater its relevance. Correlation measures the degree of association between an attribute and the label and is comprised between -1 and +1 and is obtained by Pearson's correlation. It is considered that when an attribute "a" gets a positive correlation, it is directly proportional to the label. On the other hand, when the attribute "a" gets a negative correlation, it is inversely proportional to the label. It should be noted that the "Weight by Correlation" operator cannot be applied to polynomial attributes, since in these, the weights are considered random [55].

The "k" attributes of greater weight were selected by the "Select by Weight" operator. In our experiments, the 'k' parameter was set to 10. According to the previous operator, the best 10 attributes, in descending order of relevance, were:

1. **CrippenLogP-** Atom-based calculation of LogP using Crippen's approach;
2. **VR2_Dt-** Normalized Randic-like eigenvector-based index from detour matrix;
3. **MDEN-33-** Molecular distance edge between all tertiary nitrogens;
4. **LipoaffinityIndex-** Is an "efficiency indicator";
5. **nHBint9-** Count of E-State descriptors of strength for potential Hydrogen Bonds of path length 9;
6. **MATS4v-** Moran autocorrelation - lag 4/weighted by van der Waals volumes;
7. **XLogP-** Prediction of logP based on the atom-type method called XLogP;
8. **AATSC4v-** Average centered Broto-Moreau autocorrelation - lag 4/weighted by van der Waals volumes;
9. **nF6HeteroRing-** Number of 6-membered fused rings containing heteroatoms (N, O, P, S, or halogens);
10. **GATS4v-** Geary autocorrelation - lag 4/weighted by van der Waals volumes [4];

[4] http://www.scbdd.com/padel_desc/descriptors/

The selection of attributes by this operator is important because it helps to perceive which descriptors most significantly influence the activity of the molecules. These 10 attributes were selected as the best and used for the 4 classification algorithms used in the experiment.

Finally, the algorithms were tested using the k-fold Cross-Validation methodology with k=10. This operator was automatically subdivided into the training model, which contains one of the four classification operators chosen and the test model, which includes the "Apply Model" operator and the "Performance" sub process. RapidMiner workflow of this part is shown in Figure 3.11.

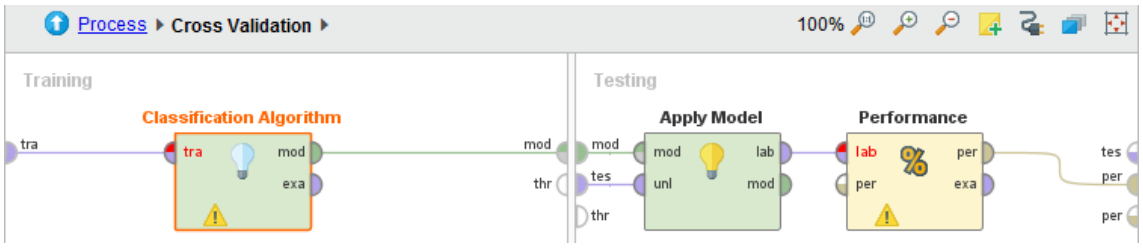

Figure 3.11: Cross-Validation process.

The "Apply Model" operator is responsible for receiving the model constructed using the training data, that is, it receives the model trained using the train data, on its first port. The test file, which contains the Drug-groupADR pairs whose rating will be predicted, is received on the second port. The last sub process, called "Performance", had as its main objective to obtain the metrics Accuracy, Recall, Precision and F-Measure (detailed in 2.2). These are presented as a percentage. The higher the percentage, the higher the performance of the model.

3.2 Case Study 2

3.2.1 Experiment 1

Data pre-processing

The present case study consists of performing the pre-processing of the data from the DrugBank data sets [5] (approached in 2.1).

This database provides a list of molecules with a lot of characteristic features. These features include the molecule id in different "standards", its SMILES representations, the family it belongs, a list of molecules with whom it is known to interact among other valuable chemical information.

Then, with the help of PaDEL tool, the SMILES representation was enriched with a set of attributes called molecular descriptors. Only 1D and 2D descriptors were generated.

Subsequently, it was necessary to convert the format of the molecules interacting with the above molecules into the SMILES format (example: 2-Methoxyestradiol - COC1=C(O)C=C2CCC3 C4CC[C@H](O)[C@@]4(C)CCC3C2=C1) To this, we resorted to the DrugBank repository. Then,

[5] https://www.drugbank.ca/about

with the help of the PaDEL Descriptors, the SMILES was enriched with a set of attributes called molecular descriptors. Only 1D and 2D descriptors were generated.

Finally, was composed of a set of data.

Chapter 4

Results

In this chapter, we present and discuss the results obtained with the realization of the experiments described in Chapter 3. Some relevant conclusions are also presented.

4.1 Case Study 1

Experiment 1

This experiment involves the use of the RapidMiner recommender extension, which provides 10 options of training algorithms for forecasting. Only 3 - Matrix Factorization, User k-NN and Slope One - were considered in this project. The objective was to measure the performance of the model by obtaining the metrics: accuracy, precision, recall and F-Measure.

The Accuracy metric represents the percentage of correct predictions made by the model. At the same time, the percentage achieved by the Precision metric is related to the actual existence of the adverse effects predicted by the system, in the real world. That is, the main error lies in the actual 0's being predicted as 1's. In turn, the Recall metric is responsible for measuring the actual adverse effects that the system can predict as actually adverse effects. Finally, the relationship obtained between the Precision and Recall metrics gives the metric F-measure. Recall that the higher F-Measure, the better the model's performance.

The following table (Table 4.1) shows the results for noise-free data:

Table 4.1: Results obtained for the metrics under study.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Matrix Factorization	45.15 (+/-1.18)	33.22 (+/-0.20)	54.05 (+/-0.57)	41.54 (+/-0.75)
User k-NN	35.54 (+/-1.35)	26.76 (+/-0.59)	50.14 (+/-4.18)	34.75 (+/-1.46)
Slope One	21.57 (+/-0.12)	20.64 (+/-0.16)	44.74 (+/-0.47)	27.89 (+/-0.23)

Observing Table 4.1, it is possible to see that the algorithm that obtained the best performance was Matrix Factorization. Note that the higher the percentage of the metrics, the better the performance of the model. The worst results were obtained with the Slope One algorithm.

Analyzing the results obtained, it can be stated that for the Matrix Factorization algorithm, the precision metric suggests that 33.22% of the adverse effects foreseen by the system are actually adverse effects in the real world. It is also observed that the recall metric indicates that the system can find 54.05% of the actual adverse effects.

Such results may be justified, since the Matrix Factorization algorithm is the only one that predicts, considering adverse effects and drugs. On the other hand, the classification of a drug made by the User k-NN algorithm is based on similarities with another group of drugs. In turn, the prediction of classifications of an adverse effect obtained with the Slope One algorithm is supported by the classifications of another adverse effect.

In the User k-NN operator, as expected, the value of k (number of users - Drug - with which the user under study will be compared) is has a strong influence on the algorithm's performance. We have observed that the performance is higher for low the values of k. We have used k=1.

The following results presented in two tables are one for each of the noise level introduced in the data set. For each noise level two data sets were generated so the algorithm was run 2 times for each noise value, and the results shown in the table correspond to the average and standard deviation of the results obtained in each one.

Table 4.2 shows the mean of the values obtained for the data with 5% of noise.

Table 4.2: Average results obtained for the metrics under study with 5% of noise.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Matrix Factorization	39.43 (+/-0.98)	30.32 (+/-0.19)	51.15 (+/-0.49)	38.74 (+/-0.67)
User k-NN	31.63 (+/-1.23)	24.24 (+/-0.55)	47.65 (+/-3.76)	32.53 (+/-1.23)
Slope One	19.53 (+/-0.17)	17.63 (+/-0.14)	42.53 (+/-0.41)	23.63 (+/-0.21)

Observing Table 4.2, it is possible to see that the algorithm that obtained the best performance was Matrix Factorization. The worst results were obtained with the Slope One algorithm.

Table 4.3 shows the mean of the values obtained for the data with 10% of noise.

Table 4.3: Average results obtained for the metrics under study with 10% of noise.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Matrix Factorization	34.75 (+/-0.87)	27.42 (+/-0.16)	47.85 (+/-0.45)	34.63 (+/-0.54)
User k-NN	28.53 (+/-1.12)	21.42 (+/-0.51)	44.63 (+/-3.01)	29.75 (+/-1.12)
Slope One	16.35 (+/-0.10)	14.65 (+/-0.11)	39.64 (+/-0.36)	19.75 (+/-0.18)

Observing Table 4.3, it is possible to see that the algorithm that obtained the best performance was Matrix Factorization. The worst results were obtained with the Slope One algorithm.

Comparing the results presented in the 3 previous tables, it is possible to state that the percentage of the values of the metrics decreases with the increase of the noise in the data, which leads to the conclusion that the noise interferes and decreases the performance of the model.

Experiment 2

Contrary to experiment 1 (described in 3.1.1), which aims to provide for an individual ADR, this focuses on predicting which groups of adverse effects are felt when taking a medicinal product.

In this experience, we have used classification algorithms that best allow us to know the most important attributes in the classification. This can be very useful and relevant insofar as such knowledge can assist the experts in understanding the cause of the adverse effect. Thus, for the accomplishment of this project, the classification algorithms chosen were: Decision tree with CART, Random Forest with CART, Naive Bayes and Support Vector Machines (LibSVM).

Then, 4 tables will be presented, one for each classification algorithm used. It should be noted that the experiment was performed for 26 adverse effects groups using a 10- fold Cross Validation. In order to optimize the results obtained, the attribute pre-selection was done with the weight by correlation.

For a better understanding of the table, it is important to note that the first column shows the ADR group to be tested. The Accuracy, Precision class=true and Recall of class=true are represented respectively in the second, third and fourth columns. Precision class=false and Recall class=false are shown in the fifth and sixth columns. Finally, in the last column, are presented the values obtained for the F-Measure metric.

In the table 4.4, it is possible to observe the results obtained using the Decision Tree algorithm with CART.

The results table show that group 8 was the one that obtained the higher Accuracy. On the other hand, the lowest Accuracy metric was obtained in group 4.

Looking of the Decision Tree generated for the best group of adverse effects under study (group 8), can see that the best molecular descriptor (root of the tree) is the CrippenLogP [1]. This descriptor means the atom-based calculation of LogP using Crippen's approach.

In the table 4.5, it is possible to observe the results obtained using the Random Forest.

Looking at the results we can see that group 8 was the one that obtained the metric Accuracy higher. On the other hand, the lowest Accuracy metric was obtained in group 9. Comparing with the previous results (shown in Table 4.4), the present Accuracy results slightly better (14/26 of the groups tested). In Figure 4.1, this improvement is perceptible.

Figure 4.1 also allows us to verify that the groups that suffered a greater difference in the Accuracy values were the groups that obtained the worse values, when using the algorithm Decision tree, that is, the groups 4, 5 and 21.

In Table 4.6, the Naive Bayes results are shown.

[1] http://www.scbdd.com/padel_desc/descriptors/

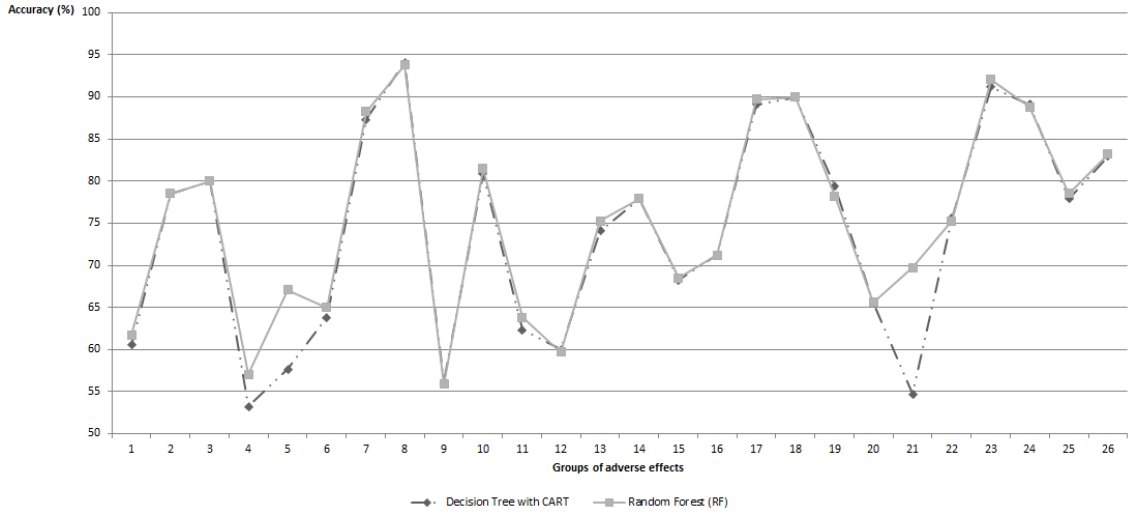

Figure 4.1: Results of the Accuracy metric obtained using the Decision Tree with CART and the Random Forest algorithms.

As with the algorithms used previously, the group that obtained a higher Accuracy was group 8. On the other hand, the lowest Accuracy was obtained in group 4. Compared with the results obtained with the Random Forest algorithm, the Naive Bayes algorithm obtained a lower Accuracy in 22/26 of the groups of adverse effects under study.

In Table 4.7, it is possible to observe the results obtained using Support Vector Machine (SVM) with kernel Radial Basis Function (RBF).

As with the algorithms used previously, the group that obtained a higher Accuracy metric was group 8. On the other hand, the lowest accuracy metric was obtained in group 4.

Figure 4.2 shows a graph with the Accuracy results obtained when using the above mentioned 4 algorithms. This allows a clearer observation of the results. Thus, it is clearly noticeable that the group that obtained the value of the highest Accuracy metric was group 8 and the best result of this is obtained when using the Support Vector Machine algorithm - about 94.41%. In general, the group that presented the worst results was group 4.

In the table 4.8, it is possible to observe the Accuracy average obtained for each algorithm used.

SVM was the algorithm that obtained the best Accuracy (about 75.32%), for the 26 groups of adverse effects under study. Then, with about 75.25%, the Random Forest algorithm appears.

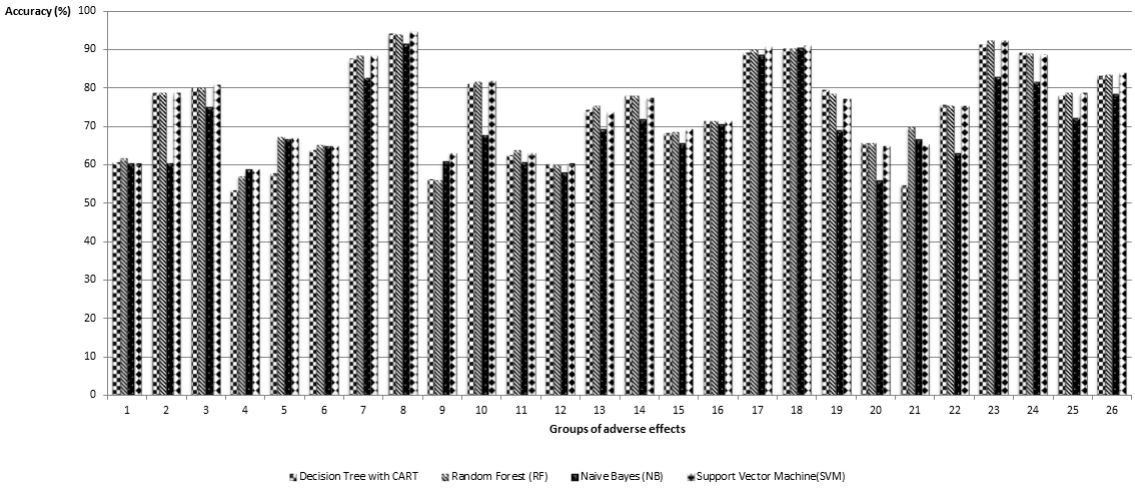

Figure 4.2: Results of the Accuracy metric obtained using the Decision Tree with CART, Random Forest, Naive Bayes and Support Vector Machine algorithms.

Table 4.4: Results obtained in the experiment 2 - Decision Trees with CART.

GroupADR	**Accuracy (%)**	**Precision (T) (%)**	**Recall (T) (%)**	**Precision (F) (%)**	**Recall (F) (%)**	**F-Measure (%)**
G1	60.59 (+/- 1.95)	61.03	97.58	44.44	3.01	5.63
G2	78.53 (+/- 1.35)	78.53	100	0.00	0.00	-
G3	80.00 (+/- 2.56)	80.84	98.54	33.33	3.03	5.56
G4	53.24 (+/- 4.64)	47.22	21.94	54.85	79.46	61.56 (+/- 15.34)
G5	57.65 (+/- 4.40)	61.90	16.88	57.05	91.40	70.05 (+/- 4.19)
G6	63.82 (+/- 3.24)	63.43	98.59	72.73	6.25	11.51
G7	87.35 (+/- 2.65)	88.13	99.00	0.00	0.00	-
G8	94.12 (+/- 2.63)	95.20	98.75	42.86	15.79	23.08
G9	56.18 (+/- 5.49)	52.01	96.27	85.71	20.11	32.58
G10	80.88 (+/- 2.71)	82.37	97.48	36.36	6.45	10.96
G11	62.35 (+/- 1.76)	62.72	99.07	0.00	0.00	-
G12	60.00 (+/- 1.44)	60.18	99.51	0.00	0.00	-
G13	74.12 (+/- 3.43)	74.85	97.60	57.14	8.89	15.38
G14	77.94 (+/- 3.54)	78.44	97.67	70.00	16.87	27.18
G15	68.24 (+/- 3.54)	68.96	98.30	20.00	0.95	1.82
G16	71.18 (+/- 3.17)	100.00	6.67	70.57	100.00	82.75 (+/- 1.82)
G17	89.12 (+/- 2.96)	90.45	98.38	0.00	0.00	-
G18	90.00 (+/- 2.70)	90.48	99.35	50.00	5.88	10.53
G19	79.41 (+/- 1.86)	79.45	98.85	78.57	14.10	23.91
G20	65.59 (+/- 3.24)	66.04	96.36	57.89	9.17	15.83
G21	54.71 (+/- 6.47)	52.69	59.76	57.14	50.00	47.41 (+/- 17.68)
G22	75.59 (+/- 2.96)	76.60	97.67	45.45	6.10	10.75
G23	91.18 (+/- 2.94)	92.49	98.40	28.57	7.41	11.76
G24	89.12 (+/- 1.35)	89.05	100.00	100.00	5.13	9.76
G25	77.94 (+/- 3.01)	50.00	13.33	79.69	96.23	87.12 (+/- 2.20)
G26	82.94 (+/- 2.20)	54.55	10.17	83.89	98.22	90.49 (+/- 1.20)

Table 4.5: Results obtained in the experiment 2 - Random Forest.

GroupADR	**Accuracy (%)**	**Precision (T) (%)**	**Recall (T) (%)**	**Precision (F) (%)**	**Recall (F) (%)**	**F-Measure (%)**
G1	61.76 (+/- 2.28)	61.99	96.14	57.89	8.27	14.47
G2	78.53 (+/- 1.35)	78.53	100	0.00	0.00	-
G3	80.00 (+/- 2.56)	80.65	98.91	25.00	1.52	2.86
G4	57.06 (+/- 6.20)	58.49	20.00	56.79	88.11	68.99 (+/- 4.70)
G5	67.06 (+/- 7.18)	70.19	47.40	65.68	83.33	73.57 (+/- 5.06)
G6	65.00 (+/- 4.82)	64.86	95.75	66.67	14.06	23.23
G7	88.24 (+/- 2.65)	88.25	100.00	0.00	0.00	-
G8	93.82 (+/- 2.06)	94.38	99.38	0.00	0.00	-
G9	55.88 (+/- 7.78)	53.90	47.20	57.29	63.69	56.95 (+/- 16.28)
G10	81.47 (+/- 1.88)	81.71	99.64	0.00	0.00	-
G11	63.82 (+/- 2.65)	64.00	97.20	60.00	7.14	12.77
G12	59.71 (+/- 2.96)	60.24	97.56	37.50	2.22	4.20
G13	75.29 (+/- 2.70)	75.30	98.80	75.00	10.00	17.65
G14	77.94 (+/- 3.54)	78.26	98.05	72.22	15.66	25.74
G15	68.53 (+/- 2.30)	69.39	97.45	40.00	3.81	6.96
G16	71.18 (+/- 3.17)	100.00	6.67	70.57	100.00	82.75 (+/- 1.82)
G17	89.71 (+/- 2.37)	90.75	98.70	20.00	3.12	5.41
G18	90.00 (+/- 2.70)	90.48	99.35	50.00	5.88	10.53
G19	78.24 (+/- 3.28)	78.66	98.47	66.67	10.26	17.78
G20	65.59 (+/- 2.65)	65.65	98.18	63.64	5.83	10.69
G21	59.71 (+/- 6.45)	57.71	59.76	61.82	57.95	57.42 (+/- 14.90)
G22	75.29 (+/- 2.35)	76.20	98.06	37.50	3.66	6.67
G23	92.09 (+/- 1.35)	92.06	100.00	0.00	0.00	-
G24	88.82 (+/- 1.76)	89.02	99.67	66.67	5.13	9.52
G25	78.53 (+/- 2.65)	62.50	6.67	78.92	98.87	87.75 (+/- 1.66)
G26	83.24 (+/- 1.35)	66.67	6.78	83.53	99.29	90.73 (+/- 0.71)

Table 4.6: Results obtained in the experiment 2 - Naive Bayes algorithm.

GroupADR	**Accuracy (%)**	**Precision (T) (%)**	**Recall (T) (%)**	**Precision (F) (%)**	**Recall (F) (%)**	**F-Measure (%)**
G1	60.29 (+/- 9.87)	70.45	59.90	49.39	60.90	54.91 (+/- 8.69)
G2	60.29 (+/- 7.81)	89.76	55.81	32.18	76.71	45.51 (+/- 6.15)
G3	74.93 (+/- 3.52)	82.03	88.32	27.27	18.46	22.02
G4	58.82 (+/- 6.58)	54.39	60.00	63.31	57.84	60.16 (+/- 7.60)
G5	66.47 (+/- 7.18)	62.05	66.88	70.69	66.13	68.49 (+/- 9.62)
G6	64.71(+/- 7.67)	69.33	77.83	53.92	42.97	47.08 (+/- 13.59)
G7	82.56 (+/- 3.77)	89.25	91.33	16.13	13.16	14.49
G8	91.43 (+/- 2.76)	95.05	95.94	13.33	11.11	12.12
G9	60.88 (+/- 11.09)	57.53	66.46	64.94	55.87	59.48 (+/- 12.86)
G10	67.65 (+/- 8.11)	88.89	69.06	30.65	61.29	41.17 (+/- 9.71)
G11	60.59 (+/- 6.34)	74.39	57.01	60.00	7.14	54.77 (+/- 10.46)
G12	57.94 (+/- 7.90)	66.46	60.98	47.37	53.33	49.82 (+/- 10.06)
G13	69.12 (+/- 3.01)	76.17	84.40	38.10	26.67	30.41 (+/- 9.63)
G14	71.76 (+/- 7.91)	78.05	87.16	37.74	24.10	29.84 (+/- 10.16)
G15	65.59 (+/- 4.93)	73.98	77.45	43.63	39.05	41.10 (+/- 6.95)
G16	70.59 (+/- 3.48)	64.71	10.48	70.90	97.45	82.02 (+/- 2.53)
G17	88.53 (+/- 3.59)	91.90	95.78	31.58	18.75	23.53
G18	90.29 (+/- 2.65)	90.99	99.02	57.14	11.76	19.51
G19	69.05 (+/- 9.36)	83.98	74.05	37.04	51.95	43.34 (+/- 14.20)
G20	55.88 (+/- 7.44)	66.36	64.55	38.10	40.00	38.13 (+/- 13.25)
G21	66.47 (+/- 9.59)	63.74	70.73	69.62	62.50	65.37 (+/- 10.97)
G22	62.94 (+/- 6.73)	78.45	70.54	29.63	39.02	33.22 (+/- 10.03)
G23	82.65 (+/- 6.37)	93.79	86.90	18.88	33.33	23.38
G24	81.47 (+/- 6.84)	88.39	91.03	10.00	5.69	8.70
G25	72.06 (+/- 5.46)	40.00	53.33	85.42	77.36	81.12 (+/- 3.91)
G26	78.24 (+/- 7.35)	40.00	50.85	89.06	83.99	86.28 (+/- 5.02)

Table 4.7: Results obtained in the experiment 2 - Support Vector Machine (SVM) algorithm.

GroupADR	**Accuracy (%)**	**Precision (T) (%)**	**Recall (T) (%)**	**Precision (F) (%)**	**Recall (F) (%)**	**F-Measure (%)**
G1	60.29 (+/- 2.71)	60.71	98.55	25.00	0.75	1.46
G2	78.53 (+/- 1.35)	78.53	100.00	0.00	0.00	-
G3	80.59 (+/- 2.35)	80.77	99.64	50.00	1.52	2.94
G4	58.82 (+/- 8.63)	58.06	34.84	59.11	78.92	67.10 (+/- 8.78)
G5	66.76 (+/- 8.11)	74.12	40.91	64.31	88.17	74.42 (+/- 5.93)
G6	64.71 (+/- 4.16)	64.56	96.23	66.67	12.50	21.05
G7	88.24 (+/- 0.00)	88.24	100.00	0.00	0.00	-
G8	94.41 (+/- 0.88)	94.41	100.00	0.00	0.00	-
G9	62.94 (+/- 10.12)	59.67	67.08	66.67	59.22	61.96 (+/- 12.79)
G10	81.76 (+/- 1.188.11)	81.76	100.00	0.00	0.00	-
G11	62.94 (+/- 1.44)	62.94	100.00	0.00	0.00	-
G12	60.29 (+/- 1.47)	60.29	100.00	0.00	0.00	-
G13	73.53 (+/- 1.32)	73.67	99.60	50.00	1.11	2.17
G14	77.35 (+/- 3.73)	77.27	99.22	80.00	9.64	17.20
G15	69.12 (+/- 1.47)	69.12	100.00	0.00	0.00	-
G16	71.18 (+/- 3.17)	100.00	6.67	70.57	100.00	82.75 (+/- 1.82)
G17	90.59 (+/- 1.18)	90.59	100.00	0.00	0.00	0.00 (+/-0.00)
G18	90.88 (+/- 3.07)	91.29	99.35	71.43	14.71	24.39
G19	77.06 (+/- 1.18)	77.06	100.00	0.00	0.00	-
G20	64.71 (+/- 0.00)	64.71	100.00	0.00	0.00	-
G21	65.29 (+/- 7.65)	61.62	74.39	70.42	56.82	62.55 (+/- 9.13)
G22	75.29 (+/- 1.95)	75.74	99.22	0.00	0.00	-
G23	92.06 (+/- 1.35)	92.06	100.00	0.00	0.00	-
G24	88.53 (+/- 0.88)	88.53	100.00	0.00	0.00	-
G25	78.53 (+/- 2.30)	100.00	2.67	78.40	100.00	87.89 (+/- 1.33)
G26	83.82 (+/- 1.47)	100.00	6.78	83.63	100.00	91.09 (+/- 0.77)

Table 4.8: Results of the Accuracy average (of the 26 adverse effects groups) for each of the algorithms used.

Algorithm	**Accuracy average (%)**
Decision Tree	73.92
Random Forest	75.25
Naive Bayes	70.43
Support Vector Machine	75.32

Chapter 5

Conclusions and Future Work

5.1 Satisfaction of Results

The occurrence of ADRs has been increasing notably in recent years and has reached the level of concern. Many studies are being carried out to predict them. In general, it has been observed that computing has proved essential in this area, since the development of predictive procedures could help to avoid many unwanted serious effects.

The main goal of this dissertation was to assess how useful Data Mining methods and algorithms could predict the adverse effect of drugs (detailed in 1.2). Thus, it was first necessary to conduct in-depth not only of the basic concepts inherent to the adverse effects of drugs, but also a set of DM methods and tools that would be very useful for predicting new adverse effects.

After the analysis of the state of the art was done, the experimental part of the project was advanced. Such as the one outlined and in order to answer the two research questions mentioned in 1.2, two case studies were addressed.

The first case study involves the realization of two experiments, whose data set was taken from the ADReCS database.

In the first experiment, were used recommendation algorithms: Matrix Factorization, Slope One and User k-NN. Considering the results obtained, it can be stated that the algorithm that obtained a model with better performance was the Matrix Factorization, since this algorithm obtained the value of the greater Accuracy metric (about 45.15%). Such results may be justified, since the Matrix Factorization algorithm is the only one that predicts, considering adverse effects and drugs. However, the results obtained were not promising.

In the second experiment, we have defined a classification tasks using only the top of the ADRs identifiers hierarchy. Thus, 26 groups of adverse effects were explored. The classification algorithms used were Decision tree with CART, Random Forest, Naive Bayes and Support Vector Machine. Each of these classification algorithms was executed 26 times, which corresponds to the number of adverse effects groups in study. It is important to remember that in this process a feature selection was made, in which the 10 best attributes were selected. Considering the results obtained, it can be stated that the algorithm that obtained a model with better performance was

the Support Vector Machine, since this algorithm obtained on average the best value of Accuracy (about 75.32%). The results obtained were satisfactory.

After analyzing the results obtained in the two experiments, it was observed that experiment 2 obtained better results. Thus, it is concluded that making individual predictions of adverse effects is quite complex, and when reducing the detail of the information, that is, when one goes up the level of hierarchy, the obtained results are better. It is important to note that this improvement of results also occurs when done with the Feature Selection, since a pre-selection of the best attributes occurs.

In response to the first research question posed in 1.2), the results obtained showed that the Data Mining tool has potential and can be really useful for the prediction of ADR's. However, there are still many points that need to be worked on and improved.

The second research question can not be fully answered because only information on 1D and 2D molecular descriptors has been used. For such, it would be interesting to use ontologies and 3D molecular descriptors.

In the course of this work, it was verified that several previous studies claim that a very relevant factor that can increase the number of ADRs episodes is the simultaneous consumption of more than one drug. Thus, it is essential to study not only the occurrence of ADR episodes in isolation, but also the possible drug interactions that may exist. These DDI may be at the origin of adverse effects that would not be proven if the drugs were taken alone.

As the results obtained in case study 1, namely for experiment 2, were satisfactory, it was found that it would be interesting to analyze DDI and to replicate the experience 2. Thus, to date, was performed the preprocessing of the data (described in 3.2.1), whose data set was taken from the DrugBank database for a possible DDI analysis.

5.2 Future Work

The simultaneous taking of several drugs is one of the most worrying and relevant issues in with regard to the occurrence of adverse effects. Thus, in a future work, it would be interesting, using the set of data obtained in 3.2.1 from the DrugBank database and replicate the process performed in experiment 2 (detailed in 3.1.2) with classification algorithms.

It may be important to include in these set of experiences the ontologies described in 2.3, to verify if additional information from the chemistry domain improves the performance of the Data Mining algorithms and, consequently, improves the performance of the model.

Finally, another proposal would be to use other operators of pre-selection of attributes, in an attempt to compare with the results already obtained, in order to analyze if it would influence the performance of the model.

References

[1] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, pages 261–273, August 2015.

[2] Yongqun He, Zuoshuang Xiang, Sirarat Sarntivijai, Luca Toldo, and Werner Ceusters. Aeo: a realism-based biomedical ontology for the representation of adverse events. *ICBO: International Conference on Biomedical Ontology*, pages 309–315, July 2011.

[3] Hossin, M., Sulaiman, and M.N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process*, 5, March 2015.

[4] Andrew M. Wilson, Lehana Thabane, and Anne Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, pages 127–134, July 2003.

[5] Mei-Chun Cai, Quan Xu, Yan-Jing Pan, Wen Pan, Nan Ji, Yin-Bo Li, Hai-Jing Jin, Ke Liu, and Zhi-Liang Ji. Adrecs: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. *Nucleic Acids Reserach*, 43:907–913, October 2014.

[6] Montiago X. LaBute, Xiaohua Zhang, Jason Lenderman, Brian J. Bennion, Sergio E. Wong, and Felice C. Lightstone. Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *OPEN ACESS*, 9, September 2014.

[7] Daniel Fábio Kawano, Leonardo Régis Leira Pereira, Julieta Mieko Ueta, and Osvaldo de Freitas. Acidentes com os medicamentos: como minimizá-los? *Revista Brasileira de Ciências Farmacêuticas*, 42:487–495, December 2006.

[8] Edouard Pauwels, Véronique Stoven, and Yoshihiro Yamanishi. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, 2011.

[9] Patrícia Mandali de Figueiredo, Alessandra Alves da Costa, Fernanda do Carmo Santa Cruz, José Romério Rabelo Melo, Marcia Santos Nogueira, and Tamara Pereira de Araújo Góes. Reações adversas a medicamentos. pages 32–39, 2009.

[10] Ming Yang, Melody Kiang, and Wei Shang. Filtering big data from social media – building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, pages 230–240, February 2015.

[11] Ricard Garcia-Serna, David Vidal, Nikita Remez, and Jordi Mestres. Large-scale predictive drug safety: From structural alerts to biological mechanisms. *American Chemical Society*, pages 1875–1887, September 2015.

[12] Organização Pan Americana da Saúde/Organização Mundial da Saúde. *A importância da farmacovigilância*. 2005.

[13] Hanlon JT, Schmader KE, Koronkowski MJ, Weinberger M, Landsman PB, Samsa GP, and Lewis IK. Adverse drug events in high risk older outpatients. 45:945–948, August 1997.

[14] Márcia Aline Leal Amorim and Marco André Cardoso. A farmacovigilância e sua importância no monitoramento das reacções adversas a medicamentos. *Revista Saúde e Desenvolvimento*, 4:33–56, December 2013.

[15] Arthur Lopes Agrizzi, Leandro Cabral Pereira, and Priscila Helena Marietto Figueira. Metodologia de busca ativa para detecção de reacções adversas a medicamentos em pacientes oncológicos. *Rev. Bras. Farm. Hosp. Ser. Saúde São Paulo*, 4:6–11, March 2013.

[16] Sérgia Maria Starling Magalhães and Wânia da Silva Carvalho. *Introduction to Data Mining and its Applications*. Atheneu edition, 2001.

[17] William E. Evans and Mary V. Relling. Pharmacogenomics: Translating functional genomics into rational therapeutics. 286:487–491, October 1999.

[18] Silvia Regina Secoli. Interacções medicamentosas: fundamentos para a prática clínica da enfermagem. 35:28–34, March 2001.

[19] Roberto DeLucia. *Farmacologia integrada: Uso Racional de Medicamentos*, volume 1. 2014.

[20] R. P. Riechelmann and A. Del Giglio. Drug interactions in oncology: how common are they? *Annals of Oncology Advance Access*, August 2009.

[21] Rogério Hoefler. Interações medicamentosas. 2010.

[22] Cristiano Moura, Francisco Acurcio, and Najara Belo. Drug-drug interactions associated with length of stay and cost of hospitalization. *J Pharm Pharmaceut Sci*, pages 266–272, September 2009.

[23] Organização Pan Americana da Saúde/Organização Mundial da Saúde. *Monitorização da Segurança de Medicamentos*. the Uppsala Monitoring Centre, 2005.

[24] M. Hauben and A. Bate. Decision support methods for the detection of adverse events in post-marketing data. *Drug Discovery Today*, 14:343–357, April 2009.

[25] European Medicines Agency. Guide on the interpretation of spontaneous case reports of suspected adverse reactions to medicines. June 2011.

[26] Rave Harpaz, William DuMouchel, Nigam H. Shah, David Madigan, Patrick Ryan, and Carol Friedman. Novel data mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther.*, June 2012.

[27] Yihui Liu and Uwe Aickelin. Feature selection in detection of adverse drug reactions from the health improvement network (thin) database. September 2014.

[28] Eugene P. van Puijenbroek, Andrew Bate, Hubert G. M. Leufkens, Marie Lindquist, Roland Orre, and Antoine C. G. Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, pages 3–10, February 2002.

[29] Antje Neubert, Harald Dormann, Hans-Ulrich Prokosch, Thomas Bürkle, Wolfgang Rascher, Reinhold Sojer, Kay Brune, and Manfred Criegee-Rieck. E-pharmacovigilance:development and implementation of a computable knowledge base to identify adverse drug reactions. *British Journal of Clinical Pharmacology*, pages 69–77, March 2013.

[30] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic Acids Research*, October 2015.

[31] Teresa Juan-Blanco, Miquel Duran-Frigola, and Patrick Aloy. Intside: a web server for the chemical and biological examination of drug side effects. *Oxford University Press*, October 2014.

[32] Allan P Davis, Cynthia G Murphy, Michael C Rosenstein, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study. *BMC Medical Genomics*, October 2008.

[33] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*, volume 11. WILEY-VCH, 2000.

[34] Berenice da Silva Junkes. Índice semi-empírico topológico: Desenvolvimento e aplicação de um novo descritor molecular em estudos de correlação quantitativa estrutura-propriedade (qspr). Technical report, Universidade Federal de Santa Catarina, Maio 2003.

[35] Mati Karelson and Victor S. Lobanov. Quantum-chemical descriptors in qsar/qspr studies. *American Chemical Society*, 96:1027–1043, Februry 1996.

[36] Andrew R.Leach and Valerie J.Gillet. *An Introduction to Chemoinformatics*. Springer, 2007.

[37] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 2011.

[38] Jie Dong, Dong-Sheng Ca, Hong-Yu Miao, Shao Liu, Bai-Chuan Deng, Yong-Huan Yun, Ning-Ning Wang, Ai-Ping Lu, Wen-Bin Zeng, and Alex F. Chen. Chemdes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journak of Cheminformatics*, 2015.

[39] Osmar R. Zaïane. Principles of knowledge discovery in databases. 1999.

[40] Michelly Martins Ferreira, Henrique Nunes Pereira Oliva, Janio Rodrigues Lopes, Mariane Luyara Campos Magalhães, Mikael Patrick Santos de Melo, Ted Ronaldo Aguiar Ramos, and Déborah Lorranny Rodrigues dos Santos. Data mining – mineração de dados uma abordagem geral. September 2014.

[41] Lissandra Luvizao Lazzarotto, Alcione de Paiva Oliveira, and Joelsio Jose Lazzarotto. Aspectos teóricos do data mining e aplicação das redes neurais em previsões de preços agropecuários. 2006.

[42] Tauller Augusto de Araújo Matos. Uma visão geral das principais tarefas de mineração de dados. *Saber Digital*, April 2012.

[43] Cássio Oliveira Camilo and João Carlos da Silva. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. August 2009.

[44] Diana Colombo Pelegrin, Diego Paz Casagrande, Merisandra Côrtes de Mattos, Priscyla Waleska Targino de Azevedo Simões, Rafael Charnovscki, and Jane Bettiol. As tarefas de associação e de classificação na shell de data mining orion. 2013.

[45] S.D.Gheware, A.S.Kejkar, and S.M.Tondare. Data mining: Task, tools, techniques and applications. *International Journal of Advanced Research in Computer and Communication Engineering*, 3:8095–8098, October 2014.

[46] S. Sumathi and S.N.Sivanandam. *Introduction to Data Mining and its Applications*, volume 29. Springer, 2006.

[47] Joyce Jackson. Data mining: A conceptual overview. *Communications of the Association for Information Systems*, 8:267–296, 2002.

[48] Daniela Cardeano. Data mining em aplicações de desenho racional de fármacos. Technical report, Faculdade de Engenharia da Universidade do Porto, July 2014.

[49] Noemi Dreyer Galvão and Heimar de Fátima Marin. Técnica de mineração de dados: uma revisão da literatura. pages 686–690, April 2009.

[50] Dr. Matthew North. *Data Mining for the Masses*. 2012.

[51] Colin Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5:13–22, 2000.

[52] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[53] Vasile Paul Bresfelean, Mihaela Bresfelean, and Ramona Lacurezeanu. Data mining tasks in a student-oriented dss. *Advanced Information Technology in Education*, pages 321–328, 2012.

[54] Michael Goebel and Le Gruenwald. A survey of data mining and knowledge discovery software tools. 1:20–33, June 1999.

[55] RapidMiner. *RapidMiner 7. Operator Reference Manual*. 2016.

[56] Yan yan Song and Ying LU. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27:130–135, 2015.

[57] Aritz Pérez, Pedro Larrañaga, and Iñaki Inza. Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, 50:342–363, 2009.

[58] Gercely da Silva and Silva. Estudo de técnicas e utilização de mineração de dados em uma base de dados da saúde pública. 2003.

[59] Cagatay Catal. Performance evaluation metrics for software fault prediction studies. *Acta Polytechnica Hungarica*, 9, 2012.

[60] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, pages 427–437, May 2009.

[61] Cristina Oprea. Performance evaluation of the data mining classification methods. *Information society and sustainable development*, pages 249–253, 2014.

[62] Payam Refaeilzadeh, Lei Tang, and Huan LIU. Cross-validation. 2009.

[63] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. 4:82–173, August 2011.

[64] Robin Burke. Hybrid recommender systems: Survey and experiments. 12:331–370, November 2002.

[65] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. 4:82–173, October 2010.

[66] WANG Panpan, Qian Qian, and LI Jingsong SHANG Zhenhong. An recommendation algorithm based on weighted slope one algorithm and user-based collaborative filtering. pages 2431–2434, 2016.

[67] Tongqiang Jiang and Wei LU. Improved slope one algorithm based on time weight. pages 2295–2297, 2013.

[68] S. Tamilmani, Dr. (Tmt), and M.Latha. An analysis of slope one based collaborative filtering algorithms for web log data. *International Journal of Multidisciplinary Research and Development*, 3:131–134, March 2016.

[69] Haipeng You, Hui Li, Yunmin Wang, and Qingzhuang Zhao. An improved collaborative filtering recommendation algorithm combining item clustering and slope one scheme. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, I, March 2015.

[70] Jingjiao Li, Limei Sun, and Jiao Wang. A slope one collaborative filtering recommendation algorithm using uncertain neighbors optimizing. pages 160–166, 2012.

[71] Zheng Wen. Recommendation system based on collaborative filtering. December 2008.

[72] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? *International Handbooks on Information Systems*, 2009.

[73] B. Chandrasekaran, John R. Josephson, and V. Richar Benjamins. What are ontologies, and why do we need them? *Journal of Biomedical Semantics*, pages 20–26, February 1999.

[74] Isabel Durán Munoz and María Rosario Bautista Zambrana. Applying ontologies to terminology: Advantages and disadvantages. *Journal of Language and Communication in Business*, pages 64–76, December 2013.

[75] Edison Andrade Martins Morais and Ana Paula L. Ambrósio. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens. December 2007.

[76] Karin Koogan Breitman and Julio Cesar Sampaio do Prado Leite. Ontologias – como e porquê criá-las. February 2015.

[77] Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, and Barry Smith. Oae: The ontology of adverse events. *Journal of Biomedical Semantics*, 2014.

[78] Paula de Matos, A Dekker, M Ennis, Janna Hastings, K Haug, S Turner, and Christoph Steinbeck. Chebi: a chemistry ontology and database. *Journal of Cheminformatics*, May 2010.

[79] Janna Hastings, Despoina Magka, Colin Batchelor, Lian Duan, Robert Stevens, Marcus Ennis, and Christoph Steinbeck. Structure-based classification and ontology in chemistry. *Journal of Cheminformatics*, 2012.

[80] Janna Hastings, Kirill Degtyarenko, Paula de Matos, Marcus Ennis, and Christoph Steinbeck. The chebi ontology: an ontology for chemistry within a biological context. 2010.

[81] Rui Camacho, Pedro Costa, Diogo Pinto, and Vítor Santos Costa. Predicting drugs adverse side-effects using a recommender-system. *Discovery Science*, pages 201–208, October 2015.